*Submitted to the Annals of Statistics*
arXiv: math.ST/1205.4645



# COVARIANCE ASSISTED SCREENING AND ESTIMATION

By Tracy Ke*,†, Jiashun Jin† and Jianqing Fan*

*Princeton University and Carnegie Mellon University*

Consider a linear model $Y = X\beta + z$, where $X = X_{n,p}$ and $z \sim N(0, I_n)$. The vector $\beta$ is unknown and it is of interest to separate its nonzero coordinates from the zero ones (i.e., variable selection). Motivated by examples in long-memory time series [11] and change-point problem [2], we are primarily interested in the case where the Gram matrix $G = X'X$ is *non-sparse* but *sparsifiable* by a finite order linear filter. We focus on the regime where signals are both *rare and weak* so that successful variable selection is very challenging but is still possible.

We approach this problem by a new procedure called the *Covariance Assisted Screening and Estimation* (CASE). CASE first uses a linear filtering to reduce the original setting to a new regression model where the corresponding Gram (covariance) matrix is sparse. The new covariance matrix induces a sparse graph, which guides us to conduct multivariate screening without visiting all the submodels. By interacting with the signal sparsity, the graph enables us to decompose the original problem into many separated small-size subproblems (if only we know where they are!). Linear filtering also induces a so-called problem of *information leakage*, which can be overcome by the newly introduced *patching* technique. Together, these give rise to CASE, which is a two-stage Screen and Clean [10, 32] procedure, where we first identify candidates of these submodels by *patching and screening*, and then re-examine each candidate to remove false positives.

For any procedure $\hat{\beta}$ for variable selection, we measure the performance by the minimax Hamming distance between the sign vectors of $\hat{\beta}$ and $\beta$. We show that in a broad class of situations where the Gram matrix is non-sparse but sparsifiable, CASE achieves the optimal rate of convergence. The results are successfully applied to a long-memory time series model and a change-point model.

*Supported in part by National Science Foundation DMS-0704337, the National Institute of General Medical Sciences of the National Institutes of Health through Grant Numbers R01GM100474 and R01-GM072611.

†Supported in part by NSF CAREER award DMS-0908613.

*AMS 2000 subject classifications:* Primary 62J05, 62J07; secondary 62C20, 62F12.

*Keywords and phrases:* Asymptotic minimaxity, Graph of Least Favorables (GOLF), Graph of Strong Dependence (GOSD), Hamming distance, multivariate screening, phase diagram, Rare and Weak signal model, sparsity, variable selection.

## 1. Introduction.

Consider a linear regression model

$$Y = X\beta + z, \qquad X = X_{n,p}, \qquad z \sim N(0, I_n). \tag{1.1}$$

The vector $\beta$ is unknown but is sparse, in the sense that only a small fraction of its coordinates is nonzero. The goal is to separate the nonzero coordinates of $\beta$ from the zero ones (i.e., variable selection).

We are primarily interested in the case where the Gram matrix $G = X'X$ is *non-sparse* but *sparsifiable*. We call $G$ *sparse* if each of its rows has relatively few 'large' elements, and we call $G$ sparsifiable if $G$ can be reduced to a sparse matrix by some simple operations (e.g. linear filtering or low-rank matrix removal). The Gram matrix plays a critical role in sparse inference, as the sufficient statistics $X'Y \sim N(G\beta, G)$. Examples where $G$ is non-sparse but sparsifiable can be found in the following application areas.

- *Change-point problem.* Recently, driven by researches on DNA copy number variation, this problem has received a resurgence of interest [18, 24, 25, 30]. While existing literature focuses on *detecting* change-points, *locating* change-points is also of major interest in many applications [1, 28, 34]. Consider a change-point model

  $$Y_i = \theta_i + z_i, \quad z_i \overset{iid}{\sim} N(0,1), \quad 1 \le i \le p, \tag{1.2}$$

  where $\theta = (\theta_1, \ldots, \theta_p)'$ is a piece-wise constant vector with jumps at relatively few locations. Let $X = X_{p,p}$ be the matrix such that $X(i,j) = 1\{j \ge i\}$, $1 \le i,j \le p$. We re-parametrize the parameters by

  $$\theta = X\beta, \qquad \text{where} \ \ \beta_k = \theta_k - \theta_{k+1}, \ \ 1 \le k \le p-1, \ \ \text{and} \ \ \beta_p = \theta_p,$$

  so that $\beta_k$ is nonzero if and only if $\theta$ has a jump at location $k$. The Gram matrix $G$ has elements $G(i,j) = \min\{i,j\}$, which is evidently non-sparse. However, adjacent rows of $G$ display a high level of similarity, and the matrix can be sparsified by a second order adjacent differencing between the rows.
- *Long-memory time series.* We consider using time-dependent data to build a prediction model for variables of interest:

  $$Y_t = \sum_j \beta_j X_{t-j} + \epsilon_t,$$

  where $\{X_t\}$ is an observed stationary time series and $\{\epsilon_t\}$ are white noise. In many applications, $\{X_t\}$ is a long-memory process. Examples include volatility process [11, 27], exchange rates, electricity demands,

and river's outflow (e.g. the Niles). Note that the problem can be reformulated as (1.1), where the Gram matrix $G = X'X$ is asymptotically close to the auto-covariance matrix of $\{X_t\}$ (say, $\Omega$). It is well-known that $\Omega$ is Toeplitz, the off-diagonal decay of which is very slow, and the matrix $L^1$-norm of which diverges as $p \to \infty$. However, the Gram matrix can be sparsified by a first order adjacent differencing between the rows.

Further examples include jump detections in (logarithm) asset prices and time series following a FARIMA model [11]. Still other examples include the factor models, where $G$ can be decomposed as the sum of a sparse matrix and a low rank (positive semi-definite) matrix. In these examples, $G$ is non-sparse, but it can be sparsified either by adjacent row differencing or low-rank matrix removal.

In this paper, motivated by the above examples, we are primarily interested in the case where $G$ is non-sparse but can be sparsified by a finite-order linear filtering. However, the idea developed in the paper applies to much broader settings, where $G$ can be sparsified by some other methods rather than linear filtering.

When $G$ is non-sparse, many existing variable selection methods face challenges. Take the lasso [5, 7, 29] for example. The success of the lasso is hinged on the so-called *irrepresentable condition* [35], which usually does not hold in the current setting as the columns of $X$ are strongly dependent. Similar conclusion can be drawn for other popular approaches, such as the SCAD [9] (despite that conditions for its success are far less stringent than those of the lasso) and the Dantzig selector [4].

In this paper, we propose a new variable selection method which we call *Covariance Assisted Screening and Estimation* (CASE). The main methodological innovation of CASE is to exploit the rich information hidden in the 'local' graphical structures among the design variables, which the lasso and many other procedures do not utilize.

In the core of CASE is *covariance assisted* multivariate screening. Screening is a well-known method of dimension reduction in Big Data. However, most literature to date has been focused on *univariate screening* or *marginal screening* [10, 15]. The major concern for extending marginal screening to (brute-force) $m$-variate screening, $m > 1$, is the computational cost. The computational complexity is at least $O(p^m)$ (excluding the complexity for obtaining $X'Y$ from $(X, Y)$; same below), which is usually unaffordable in high-dimensional problems. CASE screens only models that has $\leq m$ nodes and that form a connected subgraph of GOSD (a graph to be introduced below). As a result, in a broad context, CASE only has a computational

cost of $p$, up to a factor of multi-log$(p)$, and so overcomes the computational challenge.

We show that CASE achieves asymptotic minimaxity of Hamming distance, in the very challenging regime where the signals are both *rare and weak*; that is, only a small fraction of the coordinates of $\beta$ is nonzero, and each nonzero coordinate is relatively small. See for example [31] and the references therein. Many recent works on variable selection focus on the regime where the signals are *rare* but *strong*, and usually the *probability of exact support recovery* or the *oracle property* is used to assess the optimality of a procedure $\hat{\beta}$. When signals are both rare and weak, exact support recovery is usually impossible, and the Hamming distance—which measures the number of coordinates at which the sign vectors of $\beta$ and $\hat{\beta}$ disagree—is a more natural criterion for assessing optimality. Compared to many recent works, the theoretic framework developed in this paper is not only technically more challenging, but also scientifically more relevant.

Below, first, in Section 1.1, we introduce the Rare and Weak signal model. We then formally introduce the notion of *sparsifiability* in Section 1.2. The starting point of CASE is the use of a linear filter. In Section 1.3, we explain how linear filtering helps in variable selection by simultaneously maintaining signal sparsity and yielding the covariance matrix nearly block diagonal. In Section 1.4, we explain that linear filtering also causes a so-called problem of *information leakage*, and how to overcome such a problem by the technique of *patching*. After all these ideas are discussed, we formally introduce the CASE in Section 1.5. The computational complexity, theoretic properties, and applications of CASE are investigated in Sections 1.6-1.11.

1.1. *Rare and Weak signal model.* Our primary interest is in the situations where the signals are rare and weak, and where we have *no* information on the underlying structure of the signals. In such situations, it makes sense to use the following *Rare and Weak* signal model; see [3, 8, 20]. Fix $\epsilon \in (0,1)$ and $\tau > 0$. Let $b = (b_1, \ldots, b_p)'$ be the $p \times 1$ vector satisfying

$$b_i \stackrel{iid}{\sim} \text{Bernoulli}(\epsilon), \tag{1.3}$$

and let $\Theta_p(\tau)$ be the set of vectors

$$\Theta_p(\tau) = \{\mu \in \mathbb{R}^p : \ |\mu_i| \geq \tau, 1 \leq i \leq p\}. \tag{1.4}$$

We model $\beta$ by

$$\beta = b \circ \mu, \tag{1.5}$$

where $\mu \in \Theta_p(\tau)$ and $\circ$ is the Hadamard product (also called the coordinate-wise product). In Section 1.7, we further restrict $\mu$ to a subset of $\Theta_p(\tau)$.

In this model, $\beta_i$ is either 0 or a signal with a strength $\geq \tau$. Since we have no information on where the signals are, we assume that they appear at locations that are randomly generated. We are primarily interested in the challenging case where $\epsilon$ is small and $\tau$ is relatively small, so the signals are both rare and weak.

DEFINITION 1.1. *We call Model (1.3)-(1.5) the Rare and Weak signal model* $RW(\epsilon, \tau, \mu)$.

We remark that the theory developed in this paper is not tied to the Rare and Weak signal model, and applies to more general cases. For example, the main results can be extended to the case where we have some additional information about the underlying structure of the signals (e.g. Ising model [17]).

1.2. *Sparsifiability, linear filtering, and GOSD.* As mentioned before, we are primarily interested in the case where the Gram matrix $G$ can be sparsified by a finite-order linear filtering.

Fix an integer $h \geq 1$ and an $(h+1)$-dimensional vector $\eta = (1, \eta_1, \ldots, \eta_h)'$. Let $D = D_{h,\eta}$ be the $p \times p$ matrix satisfying

$$D_{h,\eta}(i,j) = 1\{i = j\} + \eta_1 1\{i = j-1\} + \ldots + \eta_h 1\{i = j-h\}, \quad 1 \leq i,j \leq p. \tag{1.6}$$

The matrix $D_{h,\eta}$ can be viewed as a linear operator that maps any $p \times 1$ vector $y$ to $D_{h,\eta} y$. For this reason, $D_{h,\eta}$ is also called an order $h$ linear filter [11].

For $\alpha > 0$ and $A_0 > 0$, we introduce the following class of matrices:

$$\mathcal{M}_p(\alpha, A_0) = \{\Omega \in \mathbb{R}^{p \times p} : \Omega(i,i) \leq 1, |\Omega(i,j)| \leq A_0(1+|i-j|)^{-\alpha}, 1 \leq i,j \leq p\}. \tag{1.7}$$

Matrices in $\mathcal{M}_p(\alpha, A_0)$ are not necessarily symmetric.

DEFINITION 1.2. *Fix an order $h$ linear filter $D = D_{h,\eta}$. We say that $G$ is sparsifiable by $D_{h,\eta}$ if for sufficiently large $p$, $DG \in \mathcal{M}_p(\alpha, A_0)$ for some constants $\alpha > 1$ and $A_0 > 0$.*

In the long memory time series model, $G$ can be sparsified by an order 1 linear filter. In the change-point model, $G$ can be sparsified by an order 2 linear filter.

The main benefit of linear filtering is that it induces sparsity in the Graph of Strong Dependence (GOSD) to be introduced below. Recall that the sufficient statistics $\tilde{Y} = X'Y \sim N(G\beta, G)$. Applying a linear filter $D = D_{h,\eta}$ to $\tilde{Y}$ gives

$$d \sim N(B\beta, H), \tag{1.8}$$

where $d = D(X'Y)$, $B = DG$, and $H = DGD'$. Note that no information is lost when we reduce Model (1.1) to Model (1.8).

At the same time, if $G$ is sparsifiable by $D = D_{h,\eta}$, then both the matrices $B$ and $H$ are sparse, in the sense that each row of either matrix has relatively few large coordinates. In other words, for a properly small threshold $\delta > 0$ to be determined, let $B^*$ and $H^*$ be the regularized matrices of $B$ and $H$, respectively:

$$B^*(i,j) = B(i,j)1\{|B(i,j)| \geq \delta\}, \quad H^*(i,j) = H(i,j)1\{|H(i,j)| \geq \delta\}, \quad 1 \leq i,j \leq p.$$

It is seen that

$$d \approx N(B^*\beta, H^*), \tag{1.9}$$

where each row of $B^*$ or $H^*$ has relatively few nonzeros. Compared to (1.8), (1.9) is much easier to track analytically, but it contains almost all the information about $\beta$.

The above observation naturally motivates the following graph, which we call the *Graph of Strong Dependence* (GOSD).

DEFINITION 1.3. *For a given parameter $\delta$, the GOSD is the graph $\mathcal{G}^* = (V, E)$ with nodes $V = \{1, 2, \dots, p\}$ and there is an edge between $i$ and $j$ when any of the three numbers $H^*(i,j)$, $B^*(i,j)$, and $B^*(j,i)$ is nonzero.*

DEFINITION 1.4. *A graph $\mathcal{G} = (V, E)$ is called $K$-sparse if the degree of each node $\leq K$.*

The definition of GOSD depends on a tuning parameter $\delta$, the choice of which is not critical, and it is generally sufficient if we choose $\delta = \delta_p = 1/\log(p)$; see Section 5.1 for details. With such a choice of $\delta$, it can be shown that in a general context, GOSD is $K$-sparse, where $K = K_\delta$ does not exceed a multi-$\log(p)$ term as $p \to \infty$ (see Lemma 5.1).

1.3. *Interplay between the graph sparsity and signal sparsity.* With these being said, it remains unclear how the sparsity of $\mathcal{G}^*$ helps in variable selection. In fact, even when $\mathcal{G}^*$ is 2-sparse, it is possible that a node $k$ is connected—through possible long paths—to many other nodes; it is unclear how to remove the effect of these nodes when we try to estimate $\beta_k$.

Somewhat surprisingly, the answer lies in an interesting interplay between the signal sparsity and graph sparsity. To see this point, let $S = S(\beta)$ be the support of $\beta$, and let $\mathcal{G}^*_S$ be the subgraph of $\mathcal{G}^*$ formed by the nodes in $S$ only. Given the sparsity of $\mathcal{G}^*$, if the signal vector $\beta$ is also sparse, then it is likely that the sizes of all components of $\mathcal{G}^*_S$ (a component of a graph is a maximal connected subgraph) are uniformly small. This is justified in the following lemma which is proved in [20].

LEMMA 1.1. *Suppose $\mathcal{G}^*$ is $K$-sparse and the support $S = S(\beta)$ is a realization from $\beta_j \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\pi$, where $\nu_0$ is the point mass at $0$ and $\pi$ is any distribution with support $\subseteq \mathbb{R}\backslash\{0\}$. With a probability (from randomness of $S$) at least $1 - p(e\epsilon K)^{m+1}$, $\mathcal{G}^*_S$ decomposes into many components with size no larger than $m$.*

In this paper, we are primarily interested in cases where for large $p$, $\epsilon \leq p^{-\vartheta}$ for some parameter $\vartheta \in (0, 1)$ and $K$ is bounded by a multi-$\log(p)$ term. In such cases, the decomposability of $\mathcal{G}^*_S$ holds for a finite $m$, with overwhelming probability.

Lemma 1.1 delineates an interesting picture: The set of signals decomposes into many small-size isolated signal islands (if only we know where), each of them is a component of $\mathcal{G}^*_S$, and different ones are disconnected in the GOSD. As a result, the original $p$-dimensional problem can be viewed as the aggregation of many separated small-size subproblems that can be solved parallelly. This is the key insight of this paper.

Note that the decomposability of $\mathcal{G}^*_S$ attributes to the interplay between the signal sparsity and the graph sparsity, where the latter attributes to the use of linear filtering. The decomposability is not tied to the specific model of $\beta$ in Lemma 1.1, and holds for much broader situations (e.g. when $b$ is generated by a sparse Ising model [17]).

1.4. *Information leakage and patching.* While it largely facilitates the decomposability of the model, we must note that the linear filtering also induces a so-called problem of *information leakage*. In this section, we discuss how linear filtering causes such a problem and how to overcome it by the so-called technique of *patching*.

The following notation is frequently used in this paper.

DEFINITION 1.5. *For $\mathcal{I} \subset \{1, 2, \ldots, p\}$, $\mathcal{J} \subset \{1, \cdots, N\}$, and a $p \times N$ matrix $X$, $X^{\mathcal{I}}$ denotes the $|\mathcal{I}| \times N$ sub-matrix formed by restricting the rows of $X$ to $\mathcal{I}$, and $X^{\mathcal{J},\mathcal{I}}$ denotes the $|\mathcal{J}| \times |\mathcal{I}|$ sub-matrix formed by restricting the columns of $X$ to $\mathcal{I}$ and rows to $\mathcal{J}$.*

Note that when $N = 1$, $X$ is a $p \times 1$ vector, and $X^{\mathcal{I}}$ is an $|\mathcal{I}| \times 1$ vector.

To appreciate information leakage, we first consider an idealized case where each row of $G$ has $\leq K$ nonzeros. In this case, there is no need for linear filtering, so $B = H = G$ and $d = \tilde{Y}$. Recall that $\mathcal{G}_S^*$ consists of many signal islands and let $\mathcal{I}$ be one of them. It is seen that

$$d^{\mathcal{I}} \approx N(G^{\mathcal{I},\mathcal{I}}\beta^{\mathcal{I}}, G^{\mathcal{I},\mathcal{I}}), \tag{1.10}$$

and how well we can estimate $\beta^{\mathcal{I}}$ is captured by the Fisher Information Matrix $G^{\mathcal{I},\mathcal{I}}$ [21].

Come back to the case where $G$ is non-sparse. Interestingly, despite the strong correlations, $G^{\mathcal{I},\mathcal{I}}$ continues to be the Fisher information for estimating $\beta^{\mathcal{I}}$. However, when $G$ is non-sparse, we must use a linear filtering $D = D_{h,\eta}$ as suggested, and we have

$$d^{\mathcal{I}} \approx N(B^{\mathcal{I},\mathcal{I}}\beta^{\mathcal{I}}, H^{\mathcal{I},\mathcal{I}}). \tag{1.11}$$

Moreover, letting $\mathcal{J} = \{1 \leq j \leq p :\ D(i,j) \neq 0 \text{ for some } i \in \mathcal{I}\}$, it follows that

$$B^{\mathcal{I},\mathcal{I}}\beta^{\mathcal{I}} = D^{\mathcal{I},\mathcal{J}}G^{\mathcal{J},\mathcal{I}}\beta^{\mathcal{I}}.$$

By the definition of $D$, $|\mathcal{J}| > |\mathcal{I}|$, and the dimension of the following null space $\geq 1$:

$$Null(\mathcal{I}, \mathcal{J}) = \{\xi \in \mathbb{R}^{|\mathcal{J}|} : D^{\mathcal{I},\mathcal{J}}\xi = 0\}. \tag{1.12}$$

Compare (1.11) with (1.10), and imagine the oracle situation where we are told the mean vector of $d^{\mathcal{I}}$ in both. The difference is that, we can fully recover $\beta^{\mathcal{I}}$ using (1.10), but are not able to do so with only (1.11). In other words, the information containing $\beta^{\mathcal{I}}$ is partially lost in (1.11): if we estimate $\beta^{\mathcal{I}}$ with (1.11) alone, we will never achieve the desired accuracy.

The argument is validated in Lemma 1.2 below, where the Fisher information associated with (1.11) is strictly "smaller" than $G^{\mathcal{I},\mathcal{I}}$; the difference between two matrices can be derived by taking $\mathcal{I}^+ = \mathcal{I}$ and $\mathcal{J}^+ = \mathcal{J}$ in (1.13). We call this phenomenon "information leakage".

To mitigate this, we expand the information content by including data in the neighborhood of $\mathcal{I}$. This process is called "patching". Let $\mathcal{I}^+$ be an extension of $\mathcal{I}$ by adding a few neighboring nodes, and define similarly

$\mathcal{J}^+ = \{1 \leq j \leq p : D(i,j) \neq 0 \text{ for some } i \in \mathcal{I}^+\}$ and $Null(\mathcal{I}^+, \mathcal{J}^+)$. Assuming that there is no edge between any node in $\mathcal{I}^+$ and any node in $\mathcal{G}_S^* \setminus \mathcal{I}$,

$$d^{\mathcal{I}^+} \approx N(B^{\mathcal{I}^+,\mathcal{I}}\beta^{\mathcal{I}}, H^{\mathcal{I}^+,\mathcal{I}^+}). \tag{1.13}$$

The Fisher Information Matrix for $\beta^{\mathcal{I}}$ under Model (1.13) is larger than that of (1.11), which is captured in the following lemma.

LEMMA 1.2. *The Fisher Information Matrix associated with Model (1.13) is*

$$G^{\mathcal{I},\mathcal{I}} - \left[U\big(U'\big(G^{\mathcal{J}^+,\mathcal{J}^+}\big)^{-1}U\big)^{-1}U'\right]^{\mathcal{I},\mathcal{I}}, \tag{1.14}$$

*where $U$ is any $|\mathcal{J}^+| \times (|\mathcal{J}^+| - |\mathcal{I}^+|)$ matrix whose columns form an orthonormal basis of $Null(\mathcal{I}^+, \mathcal{J}^+)$.*

When the size of $\mathcal{I}^+$ becomes appropriately large, the second matrix in (1.14) is small element-wise (and so is negligible) under mild conditions (see details in Lemma 2.3). This matrix is usually non-negligible if we set $\mathcal{I}^+ = \mathcal{I}$ and $\mathcal{J}^+ = \mathcal{J}$ (i.e., without patching).

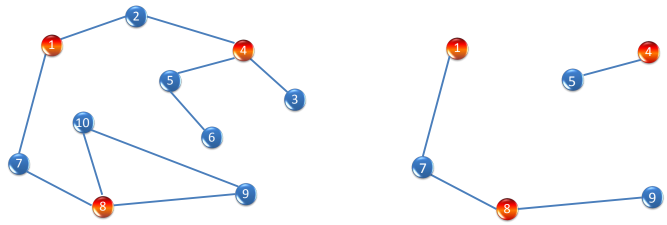
**Example 1**. We illustrate the above phenomenon with an example where $p = 5000$, $G$ is the matrix satisfying $G(i,j) = [1 + 5|i-j|]^{-0.95}$ for all $1 \leq i,j \leq p$, and $D = D_{h,\eta}$ with $h = 1$ and $\eta = (1,-1)'$. If $\mathcal{I} = \{2000\}$, then $G^{\mathcal{I},\mathcal{I}} = 1$, but the Fisher information associated with Model (1.11) is 0.5. The gap can be substantially narrowed if we patch with $\mathcal{I}^+ = \{1990, 1991, \ldots, 2010\}$, in which case the Fisher information in (1.14) is 0.904.

1.5. *Covariance Assisted Screening and Estimation (CASE).* In summary, we start from the post-filtering regression model

$$d = D\tilde{Y}, \qquad \text{where } \tilde{Y} = X'Y \text{ and } D = D_{h,\eta} \text{ is a linear filter.}$$

We have observed the following.

- *Signal Decomposability*. Linear filtering induces sparsity in GOSD, a graph constructed from the Gram matrix $G$. In this graph, the set of all true signal decomposes into many small-size signal islands, each signal island is a component of GOSD.
- *Information Patching*. Linear filtering also causes information leakage, which can be overcome by delicate patching technique.




(a) (b)

FIG 1. *Illustration of Graph of Strong Dependence (GOSD). Red: signal nodes. Blue: noise nodes. (a) GOSD with 10 nodes. (b) Nodes of GOSD that survived the PS-step.*

Naturally, these motivate a two-stage Screen and Clean variable selection approach which we call *Covariance Assisted Screening and Estimation* (CASE). CASE contains a *Patching and Screening* (PS) step, and a *Patching and Estimation* (PE) step.

- *PS-step*. We use sequential $\chi^2$-tests to identify candidates for each signal island. Each $\chi^2$-test is guided by $\mathcal{G}^*$, and aided by a carefully designed patching step. This achieves multivariate screening without visiting all submodels.
- *PE-step*. We re-investigate each candidate with penalized MLE and certain patching technique, in hope of removing false positives.

For the purpose of patching, the $PS$-step and the $PE$-step use tuning integers $\ell^{ps}$ and $\ell^{pe}$, respectively. The following notations are frequently used in this paper.

DEFINITION 1.6. *For any index $1 \leq i \leq p$, $\{i\}^{ps} = \{1 \leq j \leq p : |j - i| \leq \ell^{ps}\}$. For any subset $\mathcal{I}$ of $\{1, 2, \ldots, p\}$, $\mathcal{I}^{ps} = \cup_{i \in \mathcal{I}} \{i\}^{ps}$. Similar notation applies to $\{i\}^{pe}$ and $\mathcal{I}^{pe}$.*

We now discuss two steps in detail. Consider the $PS$-step first. Fix $m > 1$. Suppose that $\mathcal{G}^*$ has a total of $T$ connected subgraphs with size $\leq m$, which we denote by $\{\mathcal{G}_t\}_{t=1}^T$, arranged in the ascending order of the sizes, with ties breaking lexicographically.

**Example 2(a)**. We illustrate this with a toy example, where $p = 10$ and the GOSD is displayed in Figure 1(a). For $m = 3$, GOSD has $T = 30$ connected subgraphs, which we arrange as follows. Note that $\{\mathcal{G}_t\}_{t=1}^{10}$ are

singletons, $\{\mathcal{G}_t\}_{t=11}^{20}$ are connected pairs, and $\{\mathcal{G}_t\}_{t=21}^{30}$ are connected triplets.

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10\}$$
$$\{1,2\}, \{1,7\}, \{2,4\}, \{3,4\}, \{4,5\}, \{5,6\}, \{7,8\}, \{8,9\}, \{8,10\}, \{9,10\}$$
$$\{1,2,4\}, \{1,2,7\}, \{1,7,8\}, \{2,3,4\}, \{2,4,5\}, \{3,4,5\}, \{4,5,6\}, \{7,8,9\}, \{7,8,10\}, \{8,9,10\}$$

In this example, the multivariate screening exams sequentially only the 30 submodels above to decide whether any variables have additional utilities given the variables recruited before, via $\chi^2$-tests. The first 10 screening problems are just the univariate screening. After that, starting from bivariate screening, we examine the variables given those selected so far. Suppose that we are examining the variables $\{1,2\}$. The testing problem depends on how variables $\{1,2\}$ are selected in the previous steps. For example, if variables $\{1,2,4,6\}$ have already been selected in the univariate screening, there is no new recruitment and we move on to examine the submodel $\{1,7\}$. If the variables $\{1,4,6\}$ have been recruited so far, we need to test if variable $\{2\}$ has additional contributions given variable $\{1\}$. If the variables $\{4,6\}$ have been recruited in the previous steps, we will examine whether variables $\{1,2\}$ together have any significant contributions. Therefore, we have never run regression for more than two variables. Similarly, for trivariate screening, we will never run regression for more than 3 variables. Clearly, multivariate screening improves the marginal screening in that it gives significant variables chances to be recruited if it is wrongly excluded by the marginal method.

We now formally describe the procedure. The $PS$-step contains $T$ sub-stages, where we screen $\mathcal{G}_t$ sequentially, $t = 1, 2, \ldots, T$. Let $\mathcal{U}^{(t)}$ be the set of retained indices at the end of stage $t$, with $\mathcal{U}^{(0)} = \emptyset$ as the convention. For $1 \leq t \leq T$, the $t$-th sub-stage contains two sub-steps.

- (*Initial step*). Let $\hat{N} = \mathcal{U}^{(t-1)} \cap \mathcal{G}_t$ represent the set of nodes in $\mathcal{G}_t$ that have already been accepted by the end of the $(t-1)$-th sub-stage, and let $\hat{F} = \mathcal{G}_t \setminus \hat{N}$ be the set of other nodes in $\mathcal{G}_t$.
- (*Updating step*). Write for short $\mathcal{I} = \mathcal{G}_t$. Fixing a tuning parameter $\ell^{ps}$ for patching, introduce
(1.15)
$$W = (B^{\mathcal{I}^{ps},\mathcal{I}})'(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}d^{\mathcal{I}^{ps}}, \quad Q = (B^{\mathcal{I}^{ps},\mathcal{I}})'(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}(B^{\mathcal{I}^{ps},\mathcal{I}}),$$
where $W$ is a random vector and $Q$ can be thought of as the covariance matrix of $W$. Define $W_{\hat{N}}$, a subvector of $W$, and $Q_{\hat{N},\hat{N}}$, a submatrix of $Q$, as follows:
(1.16)
$$W_{\hat{N}} = (B^{\mathcal{I}^{ps},\hat{N}})'(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}d^{\mathcal{I}^{ps}}, \quad Q_{\hat{N},\hat{N}} = (B^{\mathcal{I}^{ps},\hat{N}})'(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}(B^{\mathcal{I}^{ps},\hat{N}}).$$

Introduce the test statistic

$$T(d, \hat{F}, \hat{N}) = W'Q^{-1}W - W'_{\hat{N}}(Q_{\hat{N},\hat{N}})^{-1}W_{\hat{N}}. \tag{1.17}$$

For a threshold $t = t(\hat{F}, \hat{N})$ to be determined, we update the set of retained nodes by $\mathcal{U}^{(t)} = \mathcal{U}^{(t-1)} \cup \hat{F}$ if $T(d, \hat{F}, \hat{N}) > t$, and let $\mathcal{U}^{(t)} = \mathcal{U}^{(t-1)}$ otherwise. In other words, we accept nodes in $\hat{F}$ only when they have additional utilities.

The $PS$-step terminates when $t = T$, at which point, we write $\mathcal{U}_p^* = \mathcal{U}^{(T)}$, and so

$$\mathcal{U}_p^* = \text{the set of all retained indices at the end of the } PS\text{-step}.$$

In the $PS$-step, as we screen, we accept nodes sequentially. Once a node is accepted in the $PS$-step, it stays there till the end of the $PS$-step; of course, this node could be killed in the $PE$-step. In spirit, this is similar to the well-known forward regression method, but the implementation of two methods are significantly different.

The $PS$-step uses a collection of tuning thresholds

$$\mathcal{Q} = \{t(\hat{F}, \hat{N}) : (\hat{F}, \hat{N}) \text{ are defined above}\}.$$

A convenient choice for these thresholds is to let $t(\hat{F}, \hat{N}) = 2\tilde{q}\log(p)|\hat{F}|$ for a properly small fixed constant $\tilde{q} > 0$. See Section 1.9 (and also Sections 1.10-1.11) for more discussion on the choices of $t(\hat{F}, \hat{N})$.

How does the PS-step help in variable selection? In Section 2, we show that in a broad context, provided that the tuning parameters $t(\hat{F}, \hat{N})$ are properly set, the $PS$-step has two noteworthy properties: the *Sure Screening* (SS) property and the *Separable After Screening* (SAS) property. The SS property says that $\mathcal{U}_p^*$ contains all but a negligible fraction of the true signals. The SAS property says that if we view $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^*$ (more precisely, as a subgraph of $\mathcal{G}^+$, an expanded graph of $\mathcal{G}^*$ to be introduce below), then this subgraph decomposes into many disconnected components, each having a moderate size.

Together, the SS property and the SAS property enable us to reduce the original large-scale problem to many parallel small-size regression problems, and pave the way for the $PE$-step. See Section 2 for details.

**Example 2(b)**. We illustrate the above points with the toy example in Example 2(a). Suppose after the $PS$-step, the set of retained indices $\mathcal{U}_p^*$ is $\{1, 4, 5, 7, 8, 9\}$; see Figure 1(b). In this example, we have a total of three signal nodes, $\{1\}$, $\{4\}$, and $\{8\}$, which are all retained in $\mathcal{U}_p^*$ and so the

$PS$-step yields Sure Screening. On the other hand, $\mathcal{U}_p^*$ contains a few nodes of false positives, which will be further cleaned in the $PE$-step. At the same time, viewing it as a subgraph of $\mathcal{G}^*$, $\mathcal{U}_p^*$ decomposes into two disconnected components, $\{1,7,8,9\}$ and $\{4,5\}$; compare Figure 1(a). The SS property and the SAS property enable us to reduce the original problem of 10 nodes to two parallel regression problems, one with 4 nodes, and the other with 2 nodes.

We now discuss the $PE$-step. Recall that $\ell^{pe}$ is the tuning parameter for the patching of the $PE$-step, and let $\{i\}^{pe}$ be as in Definition 1.6. The following graph can be viewed as an expanded graph of $\mathcal{G}^*$.

Definition 1.7. *Let $\mathcal{G}^+ = (V, E)$ be the graph where $V = \{1, 2, \ldots, p\}$ and there is an edge between nodes $i$ and $j$ when there exist nodes $k \in \{i\}^{pe}$ and $k' \in \{j\}^{pe}$ such that there is an edge between $k$ and $k'$ in $\mathcal{G}^*$.*

Recall that $\mathcal{U}_p^*$ is the set of retained indices at the end of the $PS$-step.

Definition 1.8. *Fix a graph $\mathcal{G}$ and its subgraph $\mathcal{I}$. We say $\mathcal{I} \trianglelefteq \mathcal{G}$ if $\mathcal{I}$ is a connected subgraph of $\mathcal{G}$, and $\mathcal{I} \lhd \mathcal{G}$ if $\mathcal{I}$ is a component (maximal connected subgraph) of $\mathcal{G}$.*

Fix $1 \leq j \leq p$. When $j \notin \mathcal{U}_p^*$, CASE estimates $\beta_j$ as 0. When $j \in \mathcal{U}_p^*$, viewing $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$, there is a unique subgraph $\mathcal{I}$ such that $j \in \mathcal{I} \lhd \mathcal{U}_p^*$. Fix two tuning parameters $u^{pe}$ and $v^{pe}$. We estimate $\beta^{\mathcal{I}}$ by minimizing

$$\min_\theta \left\{ \frac{1}{2}(d^{\mathcal{I}^{pe}} - B^{\mathcal{I}^{pe},\mathcal{I}}\theta)'(H^{\mathcal{I}^{pe},\mathcal{I}^{pe}})^{-1}(d^{\mathcal{I}^{pe}} - B^{\mathcal{I}^{pe},\mathcal{I}}\theta) + \frac{(u^{pe})^2}{2}\|\theta\|_0 \right\}, \tag{1.18}$$

where $\theta$ is an $|\mathcal{I}| \times 1$ vector where each nonzero coordinate $\geq v^{pe}$, and $\|\theta\|_0$ denotes the $L^0$-norm of $\theta$. Putting these together gives the final estimator of CASE, which we denote by $\hat{\beta}^{case} = \hat{\beta}^{case}(Y; \delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe}, D_{h,\eta}, X, p)$.

CASE uses tuning parameters $(\delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$. Earlier in this paper, we have briefly discussed how to choose $(\delta, \mathcal{Q})$. As for $m$, usually, a choice of $m = 3$ is sufficient unless the signals are relatively 'dense'. The choices of $(\ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$ are addressed in Section 1.9 (see also Sections 1.10-1.11).

1.6. *Computational complexity of CASE, comparison with multivariate screening.* The $PS$-step is closely related to the well-known method of marginal screening, and has a moderate computational complexity.

Marginal screening selects variables by thresholding the vector $d$ coordinatewise. The method is computationally fast, but it neglects 'local' graphical structures, and is thus ineffective. For this reason, in many challenging problems, it is desirable to use *multivariate screening* methods which adapt to 'local' graphical structures.

Fix $m > 1$. An $m$-variate $\chi^2$-screening procedure is one of such desired methods. The method screens all $k$-tuples of coordinates of $d$ using a $\chi^2$-test, for all $k \leq m$, in an exhaustive (brute-force) fashion. Seemingly, the method adapts to 'local' graphical structures and could be much more effective than marginal screening. However, such a procedure has a computational cost of $O(p^m)$ (excluding the computation cost for obtaining $X'Y$ from $(X, Y)$; same below) which is usually not affordable when $p$ is large.

The main computational innovation of the $PS$-step is to use a graph-assisted $m$-variate $\chi^2$-screening, which is both effective in variable selection and efficient in computation. In fact, the $PS$-step only screens $k$-tuples of coordinates of $d$ that form a connected subgraph of $\mathcal{G}^*$, for all $k \leq m$. Therefore, if $\mathcal{G}^*$ is $K$-sparse, then there are $\leq Cp(eK)^{m+1}$ connected subgraphs of $\mathcal{G}^*$ with size $\leq m$; so if $K = K_p$ is no greater than a multi-log$(p)$ term (see Definition 1.10), then the computational complexity of the $PS$-step is only $O(p)$, up to a multi-log$(p)$ term.

**Example 2(c)**. We illustrate the difference between the above three methods with the toy example in Example 2(a), where $p = 10$ and the GOSD is displayed in Figure 1(a). Suppose we choose $m = 3$. Marginal screening screens all 10 single nodes of the GOSD. The brute-force $m$-variate screening screens all $k$-tuples of indices, $1 \leq k \leq m$, with a total of $\binom{p}{1} + \ldots + \binom{p}{m} = 175$ such $k$-tuples. The $m$-variate screening in the $PS$-step only screens $k$-tuples that are connected subgraphs of $\mathcal{G}^*$, for $1 \leq k \leq m$, and in this example, we only have 30 such connected subgraphs.

The computational complexity of the $PE$-step consists two parts. The first part is the complexity of obtaining all components of $\mathcal{U}_p^*$, which is $O(pK)$ and where $K$ is the maximum degree of $\mathcal{G}^+$; note that for settings considered in this paper, $K = K_p^+$ does not exceed a multi-log$(p)$ term (see Lemma 5.2). The second part of the complexity comes from solving (1.18), which hinges on the maximal size of $\mathcal{I}$. In Lemma 2.2, we show that in a broad context, the maximal size of $\mathcal{I}$ does not exceed a constant $l_0$, provided the thresholds $\mathcal{Q}$ are properly set. Numerical studies in Section 3 also support this point. Therefore, the complexity in this part does not exceed $p \cdot 3^{l_0}$. As a result, the computational complexity of the $PE$-step is moderate. Here, the bound $O(pK + p \cdot 3^{l_0})$ is conservative; the actual computational complexity is much smaller than this.

How does CASE perform? In Sections 1.7-1.9, we set up an asymptotic framework and show that CASE is asymptotically minimax in terms of the Hamming distance over a wide class of situations. In Sections 1.10-1.11, we apply CASE to the long-memory time series and the change-point model, and elaborate the optimality of CASE in such models with the so-called phase diagram.

1.7. *Asymptotic Rare and Weak model.* In this section, we add an asymptotic framework to the Rare and Weak signal model $RW(\epsilon, \tau, \mu)$ introduced in Section 1.1. We use $p$ as the driving asymptotic parameter and tie $(\epsilon, \tau)$ to $p$ through some fixed parameters.

In particular, we fix $\vartheta \in (0, 1)$ and model the sparse parameter $\epsilon$ by

$$\epsilon = \epsilon_p = p^{-\vartheta}. \tag{1.19}$$

Note that as $p$ grows, the signal becomes increasingly sparse. At this sparsity level, it turns out that the most interesting range of signal strength is $\tau = O(\sqrt{\log(p)})$. For much smaller $\tau$, successful recovery is impossible. For much larger $\tau$, the problem is relatively easy. In light of this, we fix $r > 0$ and let

$$\tau = \tau_p = \sqrt{2r\log(p)}. \tag{1.20}$$

At the same time, recalling that in $RW(\epsilon, \tau, \mu)$, we require $\mu \in \Theta_p(\tau)$ so that $|\mu_i| \geq \tau$ for all $1 \leq i \leq p$. Fixing $a > 1$, we now further restrict $\mu$ to the following subset of $\Theta_p(\tau)$:

$$\Theta_p^*(\tau_p, a) = \{\mu \in \Theta_p(\tau_p) : \tau_p \leq |\mu_i| \leq a\tau_p,\ 1 \leq i \leq p\}. \tag{1.21}$$

DEFINITION 1.9. *We call (1.19)-(1.21) the Asymptotic Rare and Weak model $ARW(\vartheta, r, a, \mu)$.*

Requiring the strength of each signal $\leq a\tau_p$ is mainly for technical reasons, and hopefully, such a constraint can be removed in the near future. From a practical point of view, since usually we do not have sufficient information on $\mu$, we prefer to have a larger $a$: we hope that when $a$ is properly large, $\Theta_p^*(\tau_p, a)$ is broad enough, so that neither the optimal procedure nor the minimax risk needs to adapt to $a$.

Towards this end, we impose some mild regularity conditions on $a$ and the Gram matrix $G$. Let $g$ be the smallest integer such that

$$g \geq \max\{(\vartheta + r)^2/(2\vartheta r), m\}. \tag{1.22}$$

For any $p \times p$ Gram matrix $G$ and $1 \leq k \leq p$, let $\lambda_k^*(G)$ be the minimum of the smallest eigenvalues of all $k \times k$ principle sub-matrices of $G$. Introduce

$$\widetilde{\mathcal{M}}_p(c_0, g) = \{G \text{ is a } p \times p \text{ Gram matrix}, \ \lambda_k^*(G) \geq c_0, \ 1 \leq k \leq g\}. \tag{1.23}$$

For any two subsets $V_0$ and $V_1$ of $\{1, 2, \ldots, p\}$, consider the optimization problem

$$(\theta_*^{(0)}(V_0, V_1; G), \theta_*^{(1)}(V_0, V_1; G)) = \operatorname{argmin}\{(\theta^{(1)} - \theta^{(0)})' G (\theta^{(1)} - \theta^{(0)})\},$$

up to the constraints that $|\theta_i^{(k)}| \geq \tau_p$ if $i \in V_k$ and $\theta_i^{(k)} = 0$ otherwise, where $k = 0, 1$, and that in the special case of $V_0 = V_1$, the sign vectors of $\theta^{(0)}$ and $\theta^{(1)}$ are unequal. Introduce

$$a_g^*(G) = \max_{\{(V_0, V_1): |V_0 \cup V_1| \leq g\}} \max\{\|\theta_*^{(0)}(V_0, V_1; G)\|_\infty, \|\theta_*^{(1)}(V_0, V_1; G)\|_\infty\}.$$

The following lemma is elementary, so we omit the proof.

LEMMA 1.3. *For any $G \in \widetilde{\mathcal{M}}_p(c_0, g)$, there is a constant $C = C(c_0, g) > 0$ such that $a_g^*(G) \leq C$.*

In this paper, except for Section 1.11 where we discuss the change-point model, we assume

$$G \in \widetilde{\mathcal{M}}(c_0, g), \qquad a > a_g^*(G). \tag{1.24}$$

Under such conditions, $\Theta_p^*(\tau_p, a)$ is broad enough and the minimax risk (to be introduced below) does not depend on $a$. See Section 1.8 for more discussion.

For any variable selection procedure $\hat{\beta}$, we measure the performance by the Hamming distance

$$h_p(\hat{\beta}; \beta, G) = E\left[\sum_{j=1}^p 1\left\{\operatorname{sgn}(\hat{\beta}_j) \neq \operatorname{sgn}(\beta_j)\right\}\Big| X, \beta\right],$$

where the expectation is taken with respect to $\hat{\beta}$. Here, for any $p \times 1$ vector $\xi$, $\operatorname{sgn}(\xi)$ denotes the sign vector (for any number $x$, $\operatorname{sgn}(x) = 1, 0, -1$ when $x < 0$, $x = 0$, and $x > 0$ correspondingly).

Under $ARW(\vartheta, r, a, \mu)$, $\beta = b \circ \mu$, so the overall Hamming distance is

$$H_p(\hat{\beta}; \epsilon_p, \mu, G) = E_{\epsilon_p}\left[h_p(\hat{\beta}; \beta, G)\Big| X\right],$$

where $E_{\epsilon_p}$ is the expectation with respect to the law of $b$. Finally, the minimax Hamming distance under $ARW(\vartheta, r, a, \mu)$ is

$$Hamm_p^*(\vartheta, r, a, G) = \inf_{\hat{\beta}} \sup_{\mu \in \Theta_p^*(\tau_p, a)} H_p(\hat{\beta}; \epsilon_p, \mu, G).$$

In next section, we will see that the minimax Hamming distance does not depend on $a$ as long as (1.24) holds.

In many recent works, the *probability of exact support recovery* or *oracle property* is used to assess optimality, e.g. [9, 35]. However, when signals are rare and weak, exact support recovery is usually impossible, and the Hamming distance is a more appropriate criterion for assessing optimality. In comparison, study on the minimax Hamming distance is not only mathematically more demanding but also scientifically more relevant than that on the oracle property.

1.8. *Lower bound for the minimax Hamming distance.* We view the (global) Hamming distance as the aggregation of 'local' Hamming distances. To construct a lower bound for the (global) minimax Hamming distance, the key is to construct lower bounds for 'local' Hamming errors. Fix $1 \le j \le p$. The 'local' Hamming error at index $j$ is the risk we make among the neighboring indices of $j$ in GOSD, say, $\{k : d(j,k) \le g\}$, where $g$ is as in (1.22) and $d(j,k)$ is the geodesic distance between $j$ and $k$ in the GOSD. The lower bound for such a 'local' Hamming error is characterized by an exponent $\rho_j^*$, which we now introduce.

For any subset $V \subset \{1, 2, \ldots, p\}$, let $I_V$ be the $p \times 1$ vector such that the $j$-th coordinate is 1 if $j \in V$ and 0 otherwise. Fixing two subsets $V_0$ and $V_1$ of $\{1, 2, \ldots, p\}$, introduce

(1.25)
$$\varpi^*(V_0, V_1) = \tau_p^{-2}\Big(\min_{\{\theta^{(k)} = I_{V_k} \circ \mu^{(k)} : \mu^{(k)} \in \Theta_p^*(\tau_p, a), k=0,1, \mathrm{sgn}(\theta^{(0)}) \neq \mathrm{sgn}(\theta^{(1)})\}} \Big\{(\theta^{(1)} - \theta^{(0)})' G (\theta^{(1)} - \theta^{(0)})\Big\}\Big),$$

and

(1.26)
$$\rho(V_0, V_1) = \max\{|V_0|, |V_1|\}\vartheta + \frac{1}{4}\left[\left(\sqrt{\varpi^*(V_0, V_1) r} - \frac{|(|V_1| - |V_0|)|\vartheta}{\sqrt{\varpi^*(V_0, V_1) r}}\right)_+\right]^2.$$

The exponent $\rho_j^* = \rho_j^*(\vartheta, r, a, G)$ is defined by

$$\rho_j^*(\vartheta, r, a, G) = \min_{(V_0, V_1) : j \in V_0 \cup V_1} \rho(V_0, V_1). \tag{1.27}$$

The following notation $L_p$ is frequently used in this paper.

DEFINITION 1.10. *$L_p$, as a positive sequence indexed by $p$, is called a multi-$\log(p)$ term if for any fixed $\delta > 0$, $\lim_{p\to\infty} L_p p^{\delta} = \infty$ and $\lim_{p\to\infty} L_p p^{-\delta} = 0$.*

It can be shown that $L_p p^{-\rho_j^*}$ provides a lower bound for the 'local' minimax Hamming distance at index $j$, and that when (1.24) holds, $\rho_j^*(\vartheta, r, a, G)$ does not depend on $a$; see [20, Section 1.5] for details. In the remaining part of the paper, we will write it as $\rho_j^*(\vartheta, r, G)$ for short.

At the same time, in order for the aggregation of all lower bounds for 'local' Hamming errors to give a lower bound for the 'global' Hamming distance, we need to introduce *Graph of Least Favorables* (GOLF). Towards this end, recalling $g$ and $\rho(V_0, V_1)$ as in (1.22) and (1.26), respectively, let

$$(V_{0j}^*, V_{1j}^*) = \operatorname{argmin}_{\{(V_0,V_1): j\in V_0\cup V_1, |V_0\cup V_1|\le g\}} \rho(V_0, V_1),$$

and when there is a tie, pick the one that appears first lexicographically. We can think $(V_{0j}^*, V_{1j}^*)$ as the 'least favorable' configuration at index $j$; see [20, Section 1.5] for details.

DEFINITION 1.11. *GOLF is the graph $\mathcal{G}^\diamond = (V, E)$ where $V = \{1, 2, \ldots, p\}$ and there is an edge between $j$ and $k$ if and only if $(V_{0j}^* \cup V_{1j}^*) \cap (V_{0k}^* \cup V_{1k}^*) \neq \emptyset$.*

The following theorem is similar to [20, Theorem 1.1] so we omit the proof.

THEOREM 1.1. *Suppose (1.24) holds so that $\rho_j^*(\vartheta, r, a, G)$ does not depend on the parameter $a$ for sufficiently large $p$. As $p \to \infty$, $Hamm_p^*(\vartheta, r, a, G) \ge L_p[d_p(\mathcal{G}^\diamond)]^{-1} \sum_{j=1}^p p^{-\rho_j^*(\vartheta, r, G)}$, where $d_p(\mathcal{G}^\diamond)$ is the maximum degree of all nodes in $\mathcal{G}^\diamond$.*

In many examples, including those of primary interest of this paper,

$$d_p(\mathcal{G}^\diamond) \le L_p. \tag{1.28}$$

In such cases, we have the following lower bound:

$$Hamm_p^*(\vartheta, r, a, G) \ge L_p \sum_{j=1}^{p} p^{-\rho_j^*(\vartheta, r, G)}. \tag{1.29}$$

1.9. *Main results.* In this section, we show that in a broad context, provided the tuning parameters are properly set, CASE achieves the lower bound prescribed in Theorem 1.1, up to some $L_p$ terms. Therefore, the lower bound in Theorem 1.1 is tight, and CASE achieves the optimal rate of convergence.

For a given $\gamma > 0$, we focus on linear models with the Gram matrix from

$$\mathcal{M}_p^*(\gamma, g, c_0, A_1) = \widetilde{\mathcal{M}}_p(c_0, g) \cap \mathcal{M}_p(\gamma, A_1),$$

where we recall that the two terms on the right hand side are defined in (1.7) and (1.23), respectively. The following lemma is proved in Section 5.

LEMMA 1.4. *For $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$, $d_p(\mathcal{G}^\diamond) \leq L_p$. As a result,* $\mathrm{Hamm}_p^*(\vartheta, r, a, G) \geq L_p \sum_{j=1}^p p^{-\rho_j^*(\vartheta, r, G)}$.

For any linear filter $D = D_{h,\eta}$, let

$$\varphi_\eta(z) = 1 + \eta_1 z + \ldots + \eta_h z^h$$

be the so-called *characterization polynomial*. We assume the following regularity conditions.

- *Regularization Condition A (RCA).* For any root $z_0$ of $\varphi_\eta(z)$, $|z_0| \geq 1$.
- *Regularization Condition B (RCB).* There are constants $\kappa > 0$ and $c_1 > 0$ such that $\lambda_k^*(DGD') \geq c_1 k^{-\kappa}$ (see Section 1.7 for the definition of $\lambda_k^*$).

For many well-known linear filters such as adjacent differences, seasonal differences, etc., RCA is satisfied. Also, RCB is only a mild condition since $\kappa$ can be any positive number. For example, RCB holds in the change-point model and long-memory time series model with certain $D$ matrices. In general, $\kappa$ is not 0 because when $DG$ is sparse, $DGD'$ is very likely to be approximately singular and the associated value of $\lambda_k^*$ can be small when $k$ is large. This is true even for very simple $G$ (e.g. $G = I_p$, $D = D_{1,\eta}$ and $\eta = (1, -1)'$).

At the same time, these conditions can be further relaxed. For example, for the change-point problem, the Gram matrix has barely any off-diagonal decay, and does not belong to $\mathcal{M}_p^*$. Nevertheless, with slight modification in the procedure, the main results continue to hold.

CASE uses tuning parameters $(\delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$. The choice of $\delta$ is flexible, and we usually set $\delta = 1/\log(p)$. For the main theorem below, we treat $m$ as given. In practice, taking $m$ to be a small integer (say, $\leq 3$) is usually sufficient, unless the signals are relatively dense (say, $\vartheta < 1/4$). The

choice of $\ell^{ps}$ and $\ell^{pe}$ are also relatively flexible, and letting $\ell^{ps}$ be a sufficiently large constant and $\ell^{pe}$ be $(\log(p))^{\nu}$ for some constant $\nu < (1-1/\alpha)/(\kappa+1/2)$ is sufficient, where $\alpha$ is as in Definition 1.2, and $\kappa$ is as in RCB.

At the same time, in principle, the optimal choices of $(u^{pe}, v^{pe})$ are

$$u^{pe} = \sqrt{2\vartheta \log p}, \quad v^{pe} = \sqrt{2r \log p}, \tag{1.30}$$

which depend on the underlying parameters $(\vartheta, r)$ that are unknown to us. Despite this, our numeric studies in Section 3 suggest that the choices of $(u^{pe}, v^{pe})$ are relatively flexible; see Sections 3-4 for more discussions.

Last, we discuss how to choose $\mathcal{Q} = \{t(\hat{F}, \hat{N}) : (\hat{F}, \hat{N})$ are defined as in the $PS$-step$\}$. Let $t(\hat{F}, \hat{N}) = 2q\log(p)$, where $q > 0$ is a constant. It turns out that the main result (Theorem 1.2 below) holds as long as

$$q_0 \le q \le q^*(\hat{F}, \hat{N}), \tag{1.31}$$

where $q_0 > 0$ is an appropriately small constant, and for any subsets $(F, N)$,

(1.32)
$$q^*(F,N) = \max\left\{q : (|F| + |N|)\vartheta + [(\sqrt{\tilde{\omega}(F,N)r} - \sqrt{q|F|})_+]^2 \ge \psi(F,N)\right\};$$

here,
(1.33)
$$\psi(F,N) = \frac{(|F| + 2|N|)\vartheta}{2} + \begin{cases} \frac{1}{4}\omega(F,N)r, & |F| \text{ is even,} \\ \frac{\vartheta}{2} + \frac{1}{4}[(\sqrt{\omega(F,N)r} - \vartheta/\sqrt{\omega(F,N)r})_+]^2, & |F| \text{ is odd,} \end{cases}$$

with

$$\omega(F,N) = \min_{\xi \in \mathbb{R}^{|F|} : |\xi_i| \ge 1} \xi'[G^{F,F} - G^{F,N}(G^{N,N})^{-1}G^{N,F}]\xi, \tag{1.34}$$

and

$$\tilde{\omega}(F,N) = \min_{\xi \in \mathbb{R}^{|F|} : |\xi_i| \ge 1} \xi'[Q_{F,F} - Q_{F,N}(Q_{N,N})^{-1}Q_{N,F}]\xi, \tag{1.35}$$

where $Q_{F,N} = (B^{\mathcal{I}^{ps},F})'(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}(B^{\mathcal{I}^{ps},N})$ with $\mathcal{I} = F \cup N$, and $Q_{N,F}$, $Q_{F,F}$ and $Q_{N,N}$ are defined similarly. Compared to (1.15), we see that $Q_{F,N}$, $Q_{F,N}$, $Q_{N,F}$ and $Q_{N,N}$ are all submatrices of $Q$. Hence, $\tilde{\omega}(F,N)$ can be viewed as a counterpart of $\omega(F,N)$ by replacing the submatrices of $G^{\mathcal{I},\mathcal{I}}$ by the corresponding ones of $Q$.

From a practical point of view, there is a trade-off in choosing $q$: a larger $q$ would increase the number of Type II errors in the $PS$-step, but would also

reduce the computation cost in the $PE$-step. The following is a convenient choice which we recommend in this paper:

$$t(\hat{F}, \hat{N}) = 2\tilde{q}|\hat{F}|\log(p), \tag{1.36}$$

where $0 < \tilde{q} < c_0 r/4$ is a constant and $c_0$ is as in $\mathcal{M}_p^*(\gamma, g, c_0, A_1)$.

We are now ready for the main result of this paper.

THEOREM 1.2. *Suppose that for sufficiently large $p$, $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$, $D_{h,\eta}G \in \mathcal{M}_p(\alpha, A_0)$ with $\alpha > 1$, and that RCA-RCB hold. Consider $\hat{\beta}^{case} = \hat{\beta}^{case}(Y; \delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe}, D_{h,\eta}, X, p)$ with the tuning parameters specified above. Then as $p \to \infty$,*

$$\sup_{\mu \in \Theta_p^*(\tau_p, a)} H_p(\hat{\beta}^{case}; \epsilon_p, \mu, G) \leq L_p[p^{1-(m+1)\vartheta} + \sum_{j=1}^{p} p^{-\rho_j^*(\vartheta, r, G)}] + o(1). \tag{1.37}$$

Combine Lemma 1.4 and Theorem 1.2. Given the parameter $m$ is appropriately large, both the upper bound and the lower bound are tight and CASE achieves the optimal rate of convergence prescribed by

$$\text{Hamm}_p^*(\vartheta, r, a, G) = L_p \sum_{j=1}^{p} p^{-\rho_j^*(\vartheta, r, G)} + o(1). \tag{1.38}$$

Theorem 1.2 is proved in Section 2, where we explain the key idea behind the procedure, as well as the selection of the tuning parameters.

1.10. *Application to the long-memory time series model.* The long-memory time series model in Section 1 can be written as a regression model:

$$Y = X\beta + z, \qquad z \sim N(0, I_n),$$

where the Gram matrix $G$ is asymptotically Toeplitz and has slow off-diagonal decays. Without loss of generality, we consider the idealized case where $G$ is an exact Toeplitz matrix generated by a spectral density $f$:

$$G(i,j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos(|i-j|\omega) f(\omega) d\omega, \qquad 1 \leq i, j \leq p.$$

In the literature [6, 23], the spectral density for a long-memory process is usually characterized as

$$f(\omega) = |1 - e^{\sqrt{-1}\omega}|^{-2\phi} f^*(\omega), \tag{1.39}$$

where $\phi \in (0, 1/2)$ is the long-memory parameter, $f^*(\omega)$ is a positive symmetric function that is continuous on $[-\pi, \pi]$ and is twice differentiable except at $\omega = 0$.

In this model, the Gram matrix is non-sparse but it is sparsifiable. To see the point, let $\eta = (1, -1)'$ and let $D = D_{1,\eta}$ be the first-order adjacent row-differencing. On one hand, since the spectral density $f$ is singular at the origin, it follows from the Fourier analysis that

$$|G(i,j)| \geq C(1 + |i-j|)^{-(1-2\phi)}$$

and hence $G$ is non-sparse. On the other hand, it is seen that

$$B(i,j) = \sqrt{-1}\int_{|j-i|}^{|j-i|+1} \widehat{\omega f(\omega)}(\lambda)d\lambda,$$

where we recall that $B = DG$ and note that $\hat{g}$ denotes the Fourier transform of $g$. Compared to $f(\omega)$, $\omega f(\omega)$ is non-singular at the origin. Additionally, it is seen that $B \in \mathcal{M}_p(2-2\phi, A)$, where $2 - 2\phi > 1$, so $B$ is sparse (similar claim applies to $H = DGD'$). This shows that $G$ is sparsifiable by adjacent row-differencing.

In this example, there is a function $\rho^*_{lts}(\vartheta, r; f)$ that only depends on $(\vartheta, r, f)$ such that

$$\max_{\{j:\log(p)\leq j\leq p-\log(p)\}} \{|\rho^*_j(\vartheta, r, G) - \rho^*_{lts}(\vartheta, r; f)|\} \to 0, \qquad \text{as } p \to \infty,$$

where the subscript 'lts' stands for long-memory time series. The following theorem can be derived from Theorem 1.2, and is proved in Section 5.

THEOREM 1.3. *For a long-memory time series model where* $|(f^*)''(\omega)| \leq C|\omega|^{-2}$, *the minimax Hamming distance satisfies* $\mathrm{Hamm}^*_p(\vartheta, r, G) = L_p p^{1-\rho^*_{lts}(\vartheta,r;f)}$. *If we apply CASE where* $(m+1)\vartheta > \rho^*_{lts}(\vartheta, r; f)$, $\eta = (1, -1)'$, *and the tuning parameters are as in Section 1.9, then*

$$\sup_{\mu\in\Theta^*_p(\tau_p,a)} H_p(\hat{\beta}^{case}; \epsilon_p, \mu, G) \leq L_p p^{1-\rho^*_{lts}(\vartheta,r;f)} + o(1).$$

Theorem 1.3 can be interpreted by the so-called *phase diagram*. Phase diagram is a way to visualize the class of settings where the signals are so rare and weak that successful variable selection is simply impossible [19]. In detail, for a spectral density $f$ and $\vartheta \in (0, 1)$, let

$$r^*_{lts}(\vartheta) = r^*_{lts}(\vartheta; f)$$

be the unique solution of $\rho^*_{lts}(\vartheta, r; f) = 1$. Note that $r = r^*_{lts}(\vartheta)$ characterizes the minimal signal strength required for exact support recovery with high probability. We have the following proposition, which is proved in Section 5.

LEMMA 1.5. *Under the conditions of Theorem 1.3, if* $(f^*)''(0)$ *exists, then* $r^*_{lts}(\vartheta; f)$ *is a decreasing function in* $\vartheta$*, with limits* $1$ *and* $\frac{2}{\pi}\int_{-\pi}^{\pi} f^{-1}(\omega)d\omega$ *as* $\vartheta \to 1$ *and* $\vartheta \to 0$*, respectively.*

Call the two-dimensional space $\{(\vartheta, r)\colon 0 < \vartheta < 1,\, r > 0\}$ the *phase space*. Interestingly, there is a partition of the phase space as follows.

- *Region of No Recovery* $\{(\vartheta, r)\colon 0 < r < \vartheta,\, 0 < \vartheta < 1\}$. In this region, the minimax Hamming distance $\gtrsim p\epsilon_p$, where $p\epsilon_p$ is approximately the number of signals. In this region, the signals are too rare and weak and successful variable selection is impossible.
- *Region of Almost Full Recovery* $\{(\vartheta, r)\colon \vartheta < r < r^*_{lts}(\vartheta; f),\, 0 < \vartheta < 1\}$. In this region, the minimax Hamming distance is much larger than 1 but much smaller than $p\epsilon_p$. Therefore, the optimal procedure can recover most of the signals but not all of them.
- *Region of Exact Recovery* $\{(\vartheta, r)\colon r > r^*_{lts}(\vartheta; f),\, 0 < \vartheta < 1\}$. In this region, the minimax Hamming distance is $o(1)$. Therefore, the optimal procedure recovers all signals with probability $\approx 1$.

Because of the partition of the phase space, we call this the *phase diagram*.

From time to time, we wish to have a more explicit formula for the rate $\rho^*_{lts}(\vartheta, r; f)$ and the critical value $r^*_{lts}(\vartheta; f)$. In general, this is a hard problem, but both quantities can be computed numerically when $f$ is given. In Figure 2, we display the phase diagrams for the FARIMA$(0, \phi, 0)$ process where

$$f^*(\omega) = \frac{\Gamma^2(1-\phi)}{\Gamma(1-2\phi)}. \tag{1.40}$$

Take $\phi = 0.35, 0.25$ for example, $r^*_{lts}(\vartheta; f) \approx 7.14, 5.08$ for small $\vartheta$.

1.11. *Application to the change-point model.* The change-point model in the introduction can be viewed as a special case of Model (1.1), where $\beta$ is as in (1.5), and the Gram matrix satisfies

$$G(i, j) = \min\{i, j\}, \qquad 1 \le i, j \le p. \tag{1.41}$$

For technical reasons, it is more convenient *not* to normalize the diagonals of $G$ to 1.

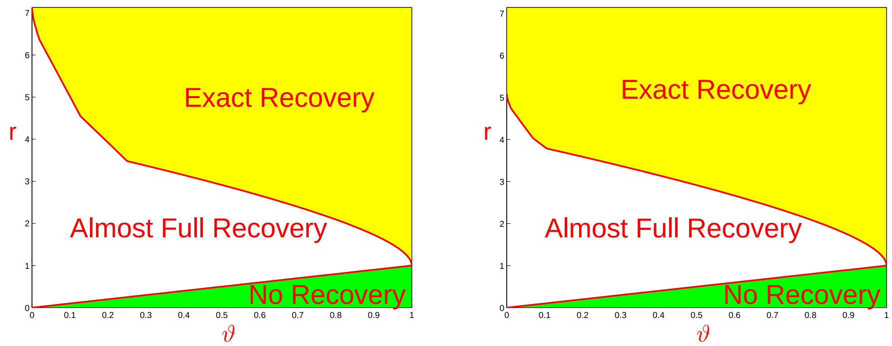



FIG 2. *Phase diagrams corresponding to the FARIMA$(0, \phi, 0)$ process. Left:* $\phi = 0.35$. *Right:* $\phi = 0.25$.

The change-point model can be viewed as an 'extreme' case of what is studied in this paper. On one hand, the Gram matrix $G$ is 'ill-posed' and each row of $G$ does not satisfy the condition of off-diagonal decay in Theorem 1.2. On the other hand, $G$ has a very special structure which can be largely exploited. In fact, if we sparsify $G$ with the linear filter $D = D_{2,\eta}$, where $\eta = (1, -2, 1)'$, it is seen that $B = DG = I_p$, and $H = DGD'$ is a tri-diagonal matrix with $H(i, j) = 2 \cdot 1\{i = j\} - 1\{|i - j| = 1\} - 1\{i = j = p\}$, which are very simple matrices. For these reasons, we modify the CASE as follows.

- Due to the simple structure of $B$, we don't need patching in the $PS$-step (i.e., $\ell^{ps} = 0$).
- For the same reason, the choices of thresholds $t(\hat{F}, \hat{N})$ are more flexible than before, and taking $t(\hat{F}, \hat{N}) = 2q \log(p)$ for a proper constant $q > 0$ works.
- However, since $H$ is 'extreme' (the smallest eigenvalue tends to 0 as $p \to \infty$), we have to modify the $PE$-step carefully.

In detail, the $PE$-step for the change-point model is as follows. Given $\ell^{pe}$, let $\mathcal{G}^+$ be as in Definition 1.7. Recall that $\mathcal{U}_p^*$ denotes the set of all retained indices at the end of the $PS$-step. Viewing $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$, and let $\mathcal{I} \lhd \mathcal{U}_p^*$ be one of its components. The goal is to split $\mathcal{I}$ into $N$ different subsets

$$\mathcal{I} = \mathcal{I}^{(1)} \cup \ldots \cup \mathcal{I}^{(N)},$$

and for each subset $\mathcal{I}^{(k)}$, $1 \leq k \leq N$, we construct a patched set $\mathcal{I}^{(k),pe}$. We then estimate $\beta^{\mathcal{I}^{(k)}}$ separately using (1.18). Putting $\beta^{\mathcal{I}^{(k)}}$ together gives our estimate of $\beta^{\mathcal{I}}$.

The subsets $\{(\mathcal{I}^{(k)}, \mathcal{I}^{(k),pe})\}_{k=1}^N$ are recursively constructed as follows. De-

note $l = |\mathcal{I}|$, $M = (\ell^{pe}/2)^{1/(l+1)}$, and write

$$\mathcal{I} = \{j_1, j_2, \cdots, j_l\}, \qquad j_1 < j_2 < \ldots < j_l.$$

First, letting $k_1$ be the largest index such that $j_{k_1} - j_{k_1 - 1} > \ell^{pe}/M$, define

$$\mathcal{I}^{(1)} = \{j_{k_1}, \cdots, j_l\}, \quad \text{and} \quad \mathcal{I}^{(1),pe} = \{j_{k_1} - \ell^{pe}/(2M), \cdots, j_l + \ell^{pe}/2\}.$$

Next, letting $k_2 < k_1$ be the largest index such that $j_{k_2} - j_{k_2 - 1} > \ell^{pe}/M^2$, define

$$\mathcal{I}^{(2)} = \{j_{k_2}, \cdots, j_{k_1}\}, \quad \mathcal{I}^{(2),pe} = \{j_{k_2} - \ell^{pe}/(2M^2), \cdots, j_{k_1} + \ell^{pe}/(2M)\}.$$

Continue this process until for some $N$, $1 \leq N \leq l$, $k_N = 1$. In this construction, for each $1 \leq k \leq N$, if we arrange all the nodes of $\mathcal{I}^{(k),pe}$ in the ascending order, then the number of nodes in front of $\mathcal{I}^{(k)}$ is significantly smaller than the number of nodes behind $\mathcal{I}^{(k)}$.

In practice, we introduce a suboptimal but much simpler patching approach as follows. Fix a component $\mathcal{I} = \{j_1, \cdots, j_l\}$ of $\mathcal{G}^+$. In this approach, instead of splitting it into smaller sets and patching them separately as in the previous approach, we patch the whole set $\mathcal{I}$ by

$$\mathcal{I}^{pe} = \{i : j_1 - \ell^{pe}/4 < i < j_l + 3\ell^{pe}/4\}, \tag{1.42}$$

and estimate $\beta^{\mathcal{I}}$ using (1.18). Our numeric studies show that two approaches have comparable performances.

Define

$$\rho^*_{cp}(\vartheta, r) = \begin{cases} \vartheta + r/4, & r/\vartheta \leq 6 + 2\sqrt{10}, \\ 3\vartheta + (r/2 - \vartheta)^2/(2r), & r/\vartheta > 6 + 2\sqrt{10}, \end{cases} \tag{1.43}$$

where 'cp' stands for change-point. Choose the tuning parameters of CASE such that

$$\ell^{pe} = 2\log(p), \quad u^{pe} = \sqrt{2\vartheta \log(p)}, \quad \text{and} \quad v^{pe} = \sqrt{2r\log(p)}, \tag{1.44}$$

that $(m+1)\vartheta \geq \rho^*_{cp}(\vartheta, r)$, and that $0 < q < \frac{r}{4}(\sqrt{2} - 1)^2$ (recall that we take $t(\hat{F}, \hat{N}) = 2q\log(p)$ for all $(\hat{F}, \hat{N})$ in the change-point setting). Note that the choice of $\ell^{pe}$ is different from that in Section 1.5. The main result in this section is the following theorem which is proved in Section 5.

THEOREM 1.4. *For the change-point model, the minimax Hamming distance satisfies* $\mathrm{Hamm}^*_p(\vartheta, r, G) = L_p p^{1 - \rho^*_{cp}(\vartheta, r)}$. *Furthermore, the CASE* $\hat{\beta}^{case}$ *with the tuning parameters specified above satisfies*

$$\sup_{\mu \in \Theta^*_p(\tau_p, a)} H_p(\hat{\beta}^{case}; \epsilon_p, \mu, G) \leq L_p p^{1 - \rho^*_{cp}(\vartheta, r)} + o(1).$$

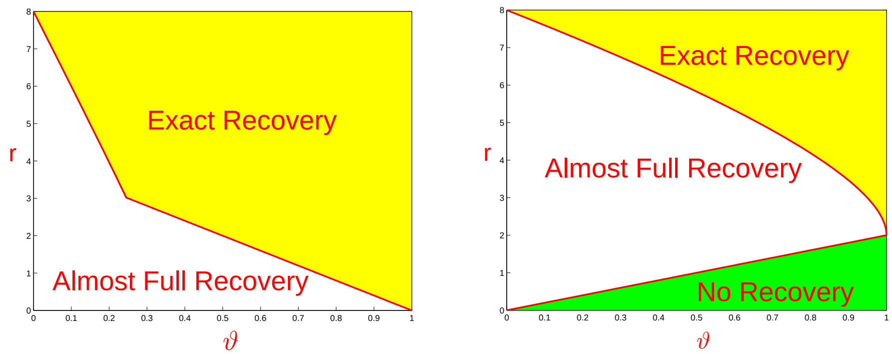



FIG 3. *Phase diagrams corresponding to the change-point model. Left: CASE; the boundary is decided by* $(4-10\vartheta)+2\sqrt{(2-5\vartheta)^2-\vartheta^2}$ *(left part) and* $4(1-\vartheta)$ *(right part). Right: hard thresholding; the upper boundary is decided by* $2(1+\sqrt{1-\vartheta})^2$ *and the lower boundary is decided by* $2\vartheta$*.*

It is noteworthy that the exponent $\rho^*_{cp}(\vartheta, r)$ has a phase change depending on the ratios of $r/\vartheta$. The insight is, when $r/\vartheta < 6+2\sqrt{10}$, the minimax Hamming distance is dominated by the Hamming errors we make in distinguishing between an isolated change point and a pair of adjacent change points, and when $r/\vartheta > 6+2\sqrt{10}$, the minimax Hamming distance is dominated by the Hamming errors of distinguishing the case of consecutive change point triplets (say, change points at $\{j-1, j, j-1\}$) from the case where we don't have a change point in the middle of the triplets (that is, the change points are only at $\{j-1, j+1\}$).

Similarly, the main results on the change-point problem can be visualized with the phase diagram, displayed in Figure 3. An interesting point is that, it is possible to have almost full recovery even when the signal strength parameter $\tau_p$ is as small as $o(\sqrt{2\log(p)})$. See the proof of Theorem 1.4 for details.

Alternatively, one may use the following approach to the change-point problem. Treat the liner change-point model as a regression model $Y = X\beta + z$ as in Section 1 (Page 2), and let $W = (X'X)^{-1}X'Y$ be the least-squares estimate. It is seen that

$$W \sim N(\beta, \Sigma),$$

where we note that $\Sigma = (X'X)^{-1}$ is tridiagonal and coincides with $H$. In this simple setting, a natural approach is to apply a coordinate-wise thresholding $\hat{\beta}_j^{thresh} = W_j 1\{|W_j| > t\}$ to locate the signals. But this neglects the covariance of $W$ in detecting the locations of the signals and is not

optimal even with the ideal choice of thresholding parameter $t_0$, since the corresponding risk satisfies

$$\sup_{\{\mu \in \Theta_p^*(\tau_p, a)\}} H_p(\hat{\beta}^{thresh}(t_0); \epsilon_p, \mu, G) = L_p p^{1-(r/2+\vartheta)^2/(2r)}.$$

The proof of this is elementary and omitted. The phase diagram of this method is displayed in Figure 3, right panel, which suggests the method is non-optimal.

Other popular methods in locating multiple change-points include the global methods (e.g. [16, 25, 29, 33]) and local methods (e.g. SaRa [24]). The global methods are usually computationally expensive and can hardly be optimal due to the strong correlation nature of this problem. Our procedure is related to the local methods but is different in important ways. Our method exploits the graphical structures and uses the GOSD to guide both the screening and cleaning, but SaRa does not utilize the graphical structures and can be shown to be non-optimal.

1.12. *Content.* The remaining sections are organized as follows. Section 2 discusses the key steps for proving Theorem 1.2. Section 3 contains numeric studies and comparisons with other methods. Section 4 contains summarizing remarks and discussions. Section 5 contains the proofs for all other theorems and lemmas in the paper.

Throughout this paper, $D = D_{h,\eta}$, $d = D(X'Y)$, $B = DG$, $H = DGD'$, and $\mathcal{G}^*$ denotes the GOSD (In contrast, $d_p$ denotes the degree of GOLF and $H_p$ denotes the Hamming distance). Also, $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real numbers and complex numbers respectively, and $\mathbb{R}^p$ denotes the $p$-dimensional real Euclidean space. Given $0 \leq q \leq \infty$, for any vector $x$, $\|x\|_q$ denotes the $L^q$-norm of $x$; for any matrix $M$, $\|M\|_q$ denotes the matrix $L^q$-norm of $M$. When $q = 2$, $\|M\|_q$ coincides with the matrix spectral norm; we shall omit the subscript $q$ in this case. When $M$ is symmetric, $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote the maximum and minimum eigenvalues of $M$ respectively. For two matrices $M_1$ and $M_2$, $M_1 \succeq M_2$ means that $M_1 - M_2$ is positive semi-definite.

## 2. Proof of the main theorem.

As mentioned before, the success of CASE relies on two noteworthy properties: the Sure Screening (SS) property and the Separable After Screening (SAS) property. In this section, we discuss the two properties in detail, and illustrate how these properties enable us to decompose the original regression problem to many small-size regression problems which can be fit separately. We then use these properties to prove Theorem 1.2.

We start with the SS property. Recall that $\mathcal{U}_p^*$ is the set of all retained indices at the end of the $PS$-step. The following lemma is proved in Section 5.

LEMMA 2.1 (SS). *Under the conditions of Theorem 1.2,*

$$\sum_{j=1}^{p} P\left(\beta_j \neq 0,\ j \notin \mathcal{U}_p^*\right) \leq L_p[p^{1-(m+1)\vartheta} + \sum_{j=1}^{p} p^{-\rho_j^*(\vartheta,r,G)}] + o(1).$$

This says that all but a negligible fraction of signals are retained in $\mathcal{U}_p^*$.

At the same time, we have the following lemma, which says that as a subgraph of $\mathcal{G}^+$, $\mathcal{U}_p^*$ splits into many disconnected components, and each component has a small size.

LEMMA 2.2 (SAS). *As $p \to \infty$, under the conditions of Theorem 1.2, there is a fixed integer $l_0 > 0$ such that with probability at least $1 - o(1/p)$, each component of $\mathcal{U}_p^*$ has a size $\leq l_0$.*

Together, these two properties enable us to decompose the original regression problem to many small-size regression problems. To see the point, let $\mathcal{I}$ be a component of $\mathcal{U}_p^*$, and $\mathcal{I}^{pe}$ be the associated patching set. Recall that $d \sim N(B\beta, H)$. If we limit our attention to nodes in $\mathcal{I}^{pe}$, then

$$d^{\mathcal{I}^{pe}} = (B\beta)^{\mathcal{I}^{pe}} + N(0, H^{\mathcal{I}^{pe},\mathcal{I}^{pe}}). \tag{2.45}$$

Denote $V = \{1, \cdots, p\} \backslash \mathcal{U}_p^*$. Write

$$(B\beta)^{\mathcal{I}^{pe}} = B^{\mathcal{I}^{pe},\mathcal{I}} \beta^{\mathcal{I}} + \xi_1 + \xi_2, \tag{2.46}$$

where

$$\xi_1 = \sum_{\mathcal{J}:\mathcal{J} \lhd \mathcal{U}_p^*, \mathcal{J} \neq \mathcal{I}} B^{\mathcal{I}^{pe},\mathcal{J}} \beta^{\mathcal{J}}, \qquad \xi_2 = B^{\mathcal{I}^{pe},V} \beta^V.$$

Now, first, by the SS property, $V$ contains only a negligible number of signals, so we expect to see that $\|\xi_2\|_\infty$ to be negligibly small. Second, by the SAS property, for any $\mathcal{J} \lhd \mathcal{U}_p^*$ and $\mathcal{J} \neq \mathcal{I}$, nodes in $\mathcal{I}$ and $\mathcal{J}$ are not connected in $\mathcal{G}^+$. By the way $\mathcal{G}^+$ is defined, it follows that nodes in $\mathcal{I}^{pe}$ and $\mathcal{J}$ are not connected in the GOSD $\mathcal{G}^*$. Therefore, we expect to see that $\|\xi_1\|_\infty$ is negligibly small as well. These heuristics are validated in the proof of Theorem 1.2; see Section 2.1 for details.

As a result,

$$d^{\mathcal{I}^{pe}} \approx N(B^{\mathcal{I}^{pe},\mathcal{I}} \beta^{\mathcal{I}}, H^{\mathcal{I}^{pe},\mathcal{I}^{pe}}), \tag{2.47}$$

where the right hand side is a small-size regression model. In other words, the original regression model decomposes into many small-size regression models, and each has a similar form to that of (2.47).

We now discuss how to fit Model (2.47). In our model $ARW(\vartheta, r, a, \mu)$, $\beta^{\mathcal{I}} = b^{\mathcal{I}} \circ \mu^{\mathcal{I}}$, and $P(\|\beta^{\mathcal{I}}\|_0 = k) \sim \epsilon_p^k$. At the same time, given a realization of $\beta^{\mathcal{I}}$, $d^{\mathcal{I}^{pe}}$ is (approximately) distributed as Gaussian as in (2.47). Combining these, for any eligible $|\mathcal{I}| \times 1$ vector $\theta$, the log-likelihood for $\beta^{\mathcal{I}} = \theta$ associated with (2.47) is

$$-\left[\frac{1}{2}(d^{\mathcal{I}^{pe}} - B^{\mathcal{I}^{pe},\mathcal{I}}\theta)'(H^{\mathcal{I}^{pe},\mathcal{I}^{pe}})^{-1}(d^{\mathcal{I}^{pe}} - B^{\mathcal{I}^{pe},\mathcal{I}}\theta) + \vartheta \log(p)\|\theta\|_0\right]. \tag{2.48}$$

Note that $\theta$ is eligible if and only if its nonzero coordinates $\geq \tau_p$ in magnitude. Comparing (2.48) with (1.18), if the tuning parameters $(u^{pe}, v^{pe})$ are set as $u^{pe} = \sqrt{2\vartheta \log(p)}$ and $v^{pe} = \sqrt{2r \log(p)}$, then the $PE$-step is actually the MLE constrained in $\Theta_p(\tau_p)$. This explains the optimality of the $PE$-step.

The last missing piece of the puzzle is how the information leakage is patched. Consider the oracle situation first where $\beta^{\mathcal{I}^c}$ is known. In such a case, by $\tilde{Y} = X'Y \sim N(G\beta, G)$, it is easy to derive that

$$\tilde{Y}^{\mathcal{I}} - G^{\mathcal{I},\mathcal{I}^c}\beta^{\mathcal{I}^c} \sim N(G^{\mathcal{I},\mathcal{I}}\beta^{\mathcal{I}}, G^{\mathcal{I},\mathcal{I}}).$$

Comparing this with Model (2.47) and applying Lemma 1.2, we see that the information leakage associated with the component $\mathcal{I}$ is captured by the matrix $[U(U'(G^{\mathcal{J}^{pe},\mathcal{J}^{pe}})^{-1}U)^{-1}U']^{\mathcal{I},\mathcal{I}}$, where $\mathcal{J}^{pe} = \{1 \leq j \leq p : D(i,j) \neq 0, \text{for some } i \in \mathcal{I}^{pe}\}$ and $U$ contains an orthonormal basis of $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$. To patch the information leakage, we have to show that this matrix has a negligible influence. This is justified in the following lemma, which is proved in Section 5.

Lemma 2.3. *(Patching). Under the conditions of Theorem 1.2, for any $\mathcal{I} \trianglelefteq \mathcal{G}^+$ such that $|\mathcal{I}| \leq l_0$, and any $|\mathcal{J}^{pe}| \times (|\mathcal{J}^{pe}| - |\mathcal{I}^{pe}|)$ matrix $U$ whose columns form an orthonormal basis of $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$,*

$$\big\|[U(U'(G^{\mathcal{J}^{pe},\mathcal{J}^{pe}})^{-1}U)^{-1}U']^{\mathcal{I},\mathcal{I}}\big\| = o(1), \qquad p \to \infty.$$

We are now ready for proving Theorem 1.2.

2.1. *Proof of Theorem 1.2 .* For short, write $\hat{\beta} = \hat{\beta}^{case}$ and $\rho_j^* = \rho_j^*(\vartheta, r, G)$. For any $\mu \in \Theta_p^*(\tau_p, a)$, write

$$H_p(\hat{\beta}; \epsilon_p, \mu, G) = I + II,$$

where
(2.49)

$$I = \sum_{j=1}^{p} P\big(\beta_j \neq 0, j \notin \mathcal{U}_p^*\big), \qquad II = \sum_{j=1}^{p} P\big(j \in \mathcal{U}_p^*,\ \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)\big).$$

Using Lemma 2.1, $I \leq L_p[p^{1-(m+1)\vartheta} + \sum_{j=1}^{p} p^{-\rho_j^*}] + o(1)$. So it is sufficient to show

$$II \leq L_p[p^{1-(m+1)\vartheta} + \sum_{j=1}^{p} p^{-\rho_j^*}] + o(1). \tag{2.50}$$

View $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$. By Lemma 2.2, there is an event $A_p$ and a fixed integer $\ell_0$ such that $P(A_p^c) \leq o(1/p)$ and that over the event $A_p$, each component of $\mathcal{U}_p^*$ has a size $\leq \ell_0$. It is seen that

$$II \leq \sum_{j=1}^{p} P\big(j \in \mathcal{U}_p^*,\ \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j),\ A_p\big) + o(1).$$

Moreover, for each $1 \leq j \leq p$, there is a unique component $\mathcal{I}$ such that $j \in \mathcal{I} \lhd \mathcal{U}_p^*$, and that $|\mathcal{I}| \leq \ell_0$ over the event $A_p$ (note that $\mathcal{I}$ depends on $\mathcal{U}_p^*$ and it is random). Since any realization of $\mathcal{I}$ must be a connected subgraph (but not necessarily a component) of $\mathcal{G}^+$,

$$II \leq \sum_{j=1}^{p} \sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0} P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j),\ A_p\big) + o(1); \tag{2.51}$$

see Definition 1.8 for the difference between $\lhd$ and $\trianglelefteq$. We stress that on the right hand side of (2.51), we have changed the meaning of $\mathcal{I}$ and use it to denote a fixed (non-random) connected subgraph of $\mathcal{G}^+$.

Next, let $\mathcal{E}(\mathcal{I}^{pe})$ be the set of nodes that are connected to $\mathcal{I}^{pe}$ by a length-1 path in $\mathcal{G}^*$:

$$\mathcal{E}(\mathcal{I}^{pe}) = \{k:\ \text{there is an edge between } k \text{ and } k' \text{ in } \mathcal{G}^* \text{ for some } k' \in \mathcal{I}^{pe}\}.$$

Heuristically, $S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe})$ is the set of signals that have major effects on $d^{\mathcal{I}^{pe}}$. Let $E_{p,\mathcal{I}}$ be the event that $(S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe})) \subset \mathcal{I}$ (note that $\mathcal{I}$ is non-random and the event is defined with respect to the randomness of $\beta$). From (2.51), we have
(2.52)

$$II \leq \sum_{j=1}^{p} \sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0} P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j),\ A_p \cap E_{p,\mathcal{I}}\big) + rem,$$

where it is seen that

$$rem \leq \sum_{j=1}^{p} \sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0} P\big(\mathcal{I} \lhd \mathcal{U}_p^*,\ A_p \cap E_{p,\mathcal{I}}^c\big). \tag{2.53}$$

The following lemma is proved in Section 5.

LEMMA 2.4. *Under the conditions of Theorem 1.2,*

$$\sum_{j=1}^{p} \sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0} P\big(\mathcal{I} \lhd \mathcal{U}_p^*, A_p \cap E_{p,\mathcal{I}}^c\big) \leq L_p \sum_{j=1}^{p} P\big(\beta_j \neq 0,\ j \notin \mathcal{U}_p^*\big).$$

Combining (2.53) with Lemma 2.4 and using Lemma 2.1,

$$rem \leq L_p[p^{1-(m+1)\vartheta} + \sum_{j=1}^{p} p^{-\rho_j^*}] + o(1). \tag{2.54}$$

Insert (2.54) into (2.52). To show (2.50), it suffices to show for each $1 \leq j \leq p$,

$$\sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0} P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j),\ A_p \cap E_{p,\mathcal{I}}\big) \leq L_p p^{-\rho_j^*}. \tag{2.55}$$

We now further reduce (2.55) to a simpler form using the sparsity of $\mathcal{G}^+$. Fix $1 \leq j \leq p$. The number of subgraphs $\mathcal{I}$ satisfying that $j \in \mathcal{I} \trianglelefteq \mathcal{G}^+$ and that $|\mathcal{I}| \leq l_0$ is no more than $C(eK_p^+)^{l_0}$ [14], where $K_p^+$ is the maximum degree of $\mathcal{G}^+$. By Lemma 5.1 and Lemma 5.2 (to be stated in Section 5), $K_p^+ \leq C(\ell^{pe})^2 K_p$, where $K_p$ is the maximum degree of $\mathcal{G}^*$, which is an $L_p$ term. Therefore, $C(eK_p^+)^{l_0}$ is also an $L_p$ term. In other words, the total number of terms in the summation of (2.55) is an $L_p$ term. As a result, to show (2.55), it suffices to show for each fixed $\mathcal{I}$ such that $j \in \mathcal{I} \trianglelefteq \mathcal{G}^+$ and $|\mathcal{I}| \leq l_0$,

$$P(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j),\ A_p \cap E_{p,\mathcal{I}}) \leq L_p p^{-\rho_j^*}. \tag{2.56}$$

Moreover, note that the left hand side of (2.56) is no more than

$$\sum_{V_0, V_1 \subset \mathcal{I}: j \in V_0 \cup V_1} P\big(\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0, \mathrm{Supp}(\hat{\beta}^{\mathcal{I}}) = V_1, \mathcal{I} \lhd \mathcal{U}_p^*, \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), A_p \cap E_{p,\mathcal{I}}\big),$$

where $V_0$ and $V_1$ are any non-random subsets satisfying the restriction. Since $|\mathcal{I}| \leq l_0$, there are only finite pairs $(V_0, V_1)$ in the summation. Therefore, to

show (2.56), it is sufficient to show for each fixed triplet $(\mathcal{I}, V_0, V_1)$ satisfying $\mathcal{I} \trianglelefteq \mathcal{G}^+$, $|\mathcal{I}| \leq l_0$, $V_0, V_1 \subset \mathcal{I}$ and $j \in V_0 \cup V_1$ that

(2.57)
$$P\big(\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0, \mathrm{Supp}(\hat{\beta}^{\mathcal{I}}) = V_1, \mathcal{I} \lhd \mathcal{U}_p^*, \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), A_p \cap E_{p,\mathcal{I}}\big) \leq L_p p^{-\rho_j^*}.$$

We now show (2.57). Fix $(\mathcal{I}, V_0, V_1)$, and write $d_1 = d^{\mathcal{I}^{pe}}$, $B_1 = B^{\mathcal{I}^{pe},\mathcal{I}}$ and $H_1 = H^{\mathcal{I}^{pe},\mathcal{I}^{pe}}$ for short. Define $\Theta_p(\mathcal{I}, a) = \{\theta \in \mathbb{R}^{|\mathcal{I}|} : \theta_j = 0 \text{ or } \tau_p \leq |\theta_j| \leq a\tau_p\}$ and $\Theta_p(\mathcal{I}) \equiv \Theta_p(\mathcal{I}, \infty)$. Since $u^{pe} = \sqrt{2\vartheta \log(p)}$ and $v^{pe} = \tau_p$, the objective function (1.18) in the $PE$-step is

$$\mathcal{L}(\theta) \equiv \frac{1}{2}(d_1 - B_1\theta)' H_1^{-1}(d_1 - B_1\theta) + \vartheta \log(p)\|\theta\|_0.$$

Over the event $\{\mathcal{I} \lhd \mathcal{U}_p^*\}$, $\hat{\beta}^{\mathcal{I}}$ minimizes the objective function, so

$$\mathcal{L}(\hat{\beta}^{\mathcal{I}}) \leq \mathcal{L}(\beta^{\mathcal{I}}).$$

As a result, the left hand side of (2.57) is no greater than

(2.58)
$$P\big(\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0, \mathrm{Supp}(\hat{\beta}^{\mathcal{I}}) = V_1, \mathcal{L}(\hat{\beta}^{\mathcal{I}}) \leq \mathcal{L}(\beta^{\mathcal{I}}), \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), A_p \cap E_{p,\mathcal{I}}\big).$$

We now calculate (2.58). Write for short $Q_1 = B_1' H_1^{-1} B_1$, $\widehat{\varpi} = \tau_p^{-2}(\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}})' Q_1 (\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}})$, and define

$$\varpi_j(V_0, V_1, \mathcal{I}) \equiv \frac{1}{\tau_p^2} \min_{(\beta^{(0)}, \beta^{(1)})} (\beta^{(1)} - \beta^{(0)})' Q_1 (\beta^{(1)} - \beta^{(0)}),$$

where the minimum is taken over $(\beta^{(0)}, \beta^{(1)})$ such that $\mathrm{sgn}(\beta_j^{(0)}) \neq \mathrm{sgn}(\beta_j^{(1)})$ and $\beta^{(k)} \in \Theta_p(\mathcal{I})$, $\mathrm{Supp}(\beta^{(k)}) = V_k$, $k = 0, 1$. Introduce

(2.59)
$$\rho_j(V_0, V_1; \mathcal{I}) = \max\{|V_0|, |V_1|\}\vartheta + \frac{1}{4}\left[\left(\sqrt{\varpi_j(V_0, V_1; \mathcal{I}) r} - \frac{|(|V_1| - |V_0|)|\vartheta}{\sqrt{\varpi_j(V_0, V_1; \mathcal{I}) r}}\right)_+\right]^2.$$

Over the event $\{\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0, \mathrm{Supp}(\hat{\beta}^{\mathcal{I}}) = V_1\}$, $\mathcal{L}(\hat{\beta}^{\mathcal{I}}) \leq \mathcal{L}(\beta^{\mathcal{I}})$ implies

(2.60)
$$-(d_1 - B_1\beta^{\mathcal{I}})' H_1^{-1} B_1 (\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}}) \geq \frac{1}{2}(\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}})' B_1' H_1^{-1} B_1 (\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}}) + (|V_1| - |V_0|)\vartheta \log(p).$$

With the notation $\widehat{\varpi}$, the right hand side of (2.60) is equal to

$$\frac{1}{2}\widehat{\varpi}\tau_p^2 + (|V_1| - |V_0|)\vartheta \log(p). \tag{2.61}$$

To simplify the left hand side of (2.60), we need the following lemma, which is proved in Section 5.

LEMMA 2.5. *For any fixed $\mathcal{I}$ such that $|\mathcal{I}| \leq l_0$, and any realization of $\beta$ over the event $E_{p,\mathcal{I}}$,*

$$(B\beta)^{\mathcal{I}^{pe}} = \zeta + B^{\mathcal{I}^{pe},\mathcal{I}}\beta^{\mathcal{I}},$$

*for some $\zeta$ satisfying $\|\zeta\| \leq C(\ell^{pe})^{1/2}[\log(p)]^{-(1-1/\alpha)}\tau_p$.*

Using Lemma 2.5, we can write $d_1 - B_1\beta^{\mathcal{I}} = \zeta + H_1^{1/2}\tilde{z}$, where $\zeta$ is as in Lemma 2.5 and $\tilde{z} \sim N(0, I_{|\mathcal{I}^{pe}|})$. It follows that the left hand side of (2.60) is equal to

$$-\zeta' H_1^{-1}B_1(\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}}) + \tilde{z}' H_1^{-1/2}B_1(\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}}). \tag{2.62}$$

First, by Cauchy-Schwartz inequality, the second term in (2.62) is no larger than $\|\tilde{z}\|\sqrt{\widehat{\varpi}\tau_p^2}$. Second, we argue that the first term in (2.62) is $o(\|\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}}\|\tau_p)$. To see the point, it suffices to check $\|B_1'H_1^{-1}\zeta\| = o(\tau_p)$. In fact, note that since $B \in \mathcal{M}_p(\alpha, A_0)$, $\|B_1\| \leq \|B\| \leq C$; in addition, by RCB, $\|H_1^{-1}\| \leq c_1^{-1}|\mathcal{I}^{pe}|^{\kappa} = O((\ell^{pe})^{\kappa})$. Applying Lemma 2.5 and noticing that $\ell^{pe} = (\log(p))^{\nu}$ with $\nu < (1-1/\alpha)/(\kappa+1/2)$, we have $\|B_1'H_1^{-1}\zeta\| \leq \|B_1\|\|H_1^{-1}\|\|\zeta\| \leq C(\ell^{pe})^{\kappa+1/2}[\log(p)]^{-(1-1/\alpha)}\tau_p$, and the claim follows. Third, from Lemma 1.2 and Lemma 2.3, $\|G^{\mathcal{I},\mathcal{I}} - Q_1\| = o(1)$ as $p$ grows. So for sufficiently large $p$, $\lambda_{\min}(Q_1) \geq \frac{1}{2}\lambda_{\min}(G^{\mathcal{I},\mathcal{I}}) \geq C$ for some constant $C > 0$. It follows from the definition of $\widehat{\varpi}$ that $\sqrt{\widehat{\varpi}\tau_p^2} \geq C\|\hat{\beta}^{\mathcal{I}} - \beta^{\mathcal{I}}\|$. Combining these with (2.62), over the event $A_p \cap E_{p,\mathcal{I}}$, the left hand side of (2.60) is no larger than

$$\sqrt{\widehat{\varpi}\tau_p^2}\ \left(\|\tilde{z}\| + o(\tau_p)\right). \tag{2.63}$$

Inserting (2.61) and (2.63) into (2.60), we see that over the event $\{\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0, \mathrm{Supp}(\hat{\beta}^{\mathcal{I}}) = V_1, \mathcal{L}(\hat{\beta}^{\mathcal{I}}) \leq \mathcal{L}(\beta^{\mathcal{I}}), A_p \cap E_{p,\mathcal{I}}\}$,

$$\|\tilde{z}\| \geq \frac{1}{2}\left(\sqrt{\widehat{\varpi}r} + \frac{(|V_1| - |V_0|)\vartheta}{\sqrt{\widehat{\varpi}r}}\right)_+ \sqrt{2\log(p)} + o(\sqrt{\log(p)}). \tag{2.64}$$

Introduce two functions defined over $(0, \infty)$: $J_1(x) = |V_0|\vartheta + \frac{1}{4}\big[\big(\sqrt{x} + \frac{(|V_1|-|V_0|)\vartheta}{\sqrt{x}}\big)_+\big]^2$ and $J_2(x) = \max\{|V_0|, |V_1|\}\vartheta + \frac{1}{4}\big[\big(\sqrt{x} - \frac{|(|V_1|-|V_0|)|\vartheta}{\sqrt{x}}\big)_+\big]^2$. By elementary calculations, $J_1(x) \geq J_2(y)$ for any $x \geq y > 0$. Now, by these notations, (2.64) can be written equivalently as $\|\tilde{z}\|^2 \geq [J_1(\widehat{\varpi}r) - |V_0|\vartheta] \cdot 2\log(p) + o(\log(p))$, and $\rho_j(V_0, V_1; \mathcal{I})$ defined in (2.59) reduces to $J_2(\varpi_j r)$, where $\varpi_j = \varpi_j(V_0, V_1; \mathcal{I})$ for short. Moreover, when $\mathrm{sgn}(\hat{\beta}_j^{\mathcal{I}}) \neq \mathrm{sgn}(\beta_j^{\mathcal{I}})$, $\widehat{\varpi} \geq \varpi_j$ by definition, and hence $J_1(\widehat{\varpi}r) \geq J_2(\varpi_j r)$. Combining these,

it follows from (2.64) that over the event $\{\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0, \mathrm{Supp}(\hat{\beta}^{\mathcal{I}}) = V_1, \mathcal{L}(\hat{\beta}^{\mathcal{I}}) \leq \mathcal{L}(\beta^{\mathcal{I}}), \mathrm{sgn}(\hat{\beta}_j^{\mathcal{I}}) \neq \mathrm{sgn}(\beta_j^{\mathcal{I}}), A_p \cap E_{p,\mathcal{I}}\}$,

$$\|\tilde{z}\|^2 \geq \big[\rho_j(V_0, V_1; \mathcal{I}) - |V_0|\vartheta\big] \cdot 2\log(p) + o(\log(p)),$$

where compared to (2.64), the right hand side is now non-random. It follows that the probability in (2.58)

$$\leq P\bigg(\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0,\ \|\tilde{z}\|^2 \geq \big[\rho_j(V_0, V_1; \mathcal{I}) - |V_0|\vartheta\big] \cdot 2\log(p) + o(\log(p))\bigg). \tag{2.65}$$

Recall that $\beta^{\mathcal{I}} = b^{\mathcal{I}} \circ \mu^{\mathcal{I}}$, where $b_j$'s are independent Bernoulli variables with surviving probability $\epsilon_p = p^{-\vartheta}$. It follows that $P(\mathrm{Supp}(\beta^{\mathcal{I}}) = V_0) = L_p p^{-|V_0|\vartheta}$. Moreover, $\|\tilde{z}\|^2$ is independent of $\beta^{\mathcal{I}}$, and is distributed as $\chi^2$ with degree of freedom $|\mathcal{I}^{ps}| \leq L_p$. From basic properties of the $\chi^2$-distribution, $P(\|\tilde{z}\|^2 > 2C\log(p) + o(\log(p))) \leq L_p p^{-C}$ for any $C > 0$. Combining these, we find that the term in (2.65)

$$\leq L_p p^{-|V_0|\vartheta - [\rho_j(V_0,V_1;\mathcal{I}) - |V_0|\vartheta]} = L_p p^{-\rho_j(V_0,V_1;\mathcal{I})}. \tag{2.66}$$

The claim follows by combining (2.66) and the following lemma.

LEMMA 2.6. *Under conditions of Theorem 1.2, for any $(j, V_0, V_1, \mathcal{I})$ satisfying $\mathcal{I} \trianglelefteq \mathcal{G}^+$, $|\mathcal{I}| \leq l_0$, $V_0, V_1 \subset \mathcal{I}$ and $j \in V_0 \cup V_1$,*

$$\rho_j(V_0, V_1; \mathcal{I}) \geq \rho_j^*(\vartheta, r, G) + o(1).$$

□

**3. Simulations.** We conducted a small-scale simulation experiment. The goal is to investigate how CASE performs with representative parameters. We focus the study on the change-point model and long-memory time series model discussed earlier.

3.1. *Change-point model.* In this section, we use Model (1.2) to investigate the performance of CASE in identifying multiple change-points. For a given set of parameters $(p, \vartheta, r, a)$, we set $\epsilon_p = p^{-\vartheta}$ and $\tau_p = \sqrt{2r\log(p)}$. First, we generate a $(p-1) \times 1$ vector $\beta$ by $\beta_j \stackrel{iid}{\sim} (1-\epsilon_p)\nu_0 + \frac{\epsilon_p}{2} U(\tau_p, a\tau_p) + \frac{\epsilon_p}{2} U(-a\tau_p, -\tau_p)$, where $U(s,t)$ is the uniform distribution over $[s,t]$ (when $s = t$, $U(s,t)$ represents the point mass at $s$). Next, we construct the mean vector $\theta$ in Model (1.2) by $\theta_j = \theta_{j-1} + \beta_{j-1}$, $2 \leq j \leq p$. Last, we generate the data vector $Y$ by $Y \sim N(\theta, I_p)$.

CASE uses tuning parameters $(\delta, m, \mathcal{Q}, \ell^{pe}, u^{pe}, v^{pe})$. Among these tuning parameters, $(\delta, m, \mathcal{Q}, \ell^{pe})$ are reasonably flexible. The optimal choice of $(u^{pe}, v^{pe})$ depends on the unknown parameters $(\epsilon_p, \tau_p)$, and how to estimate them in general settings is a lasting open problem (even for linear models with orthogonal designs). However, we note that, first, Experiment 1.1(a) shows that if we mis-specify $(\epsilon_p, \tau_p)$ by a reasonably small amount and use them to decide the optimal choice of $(u^{pe}, v^{pe})$, then the misspecification usually has only a negligible effect on the performance of CASE. Second, in some cases, $(\epsilon_p, \tau_p)$ can be estimated satisfactorily; see Experiment 1.1(b). For these reasons, in most experiments below, we set the tuning parameters in a way by assuming $(\epsilon_p, \tau_p)$ (or equivalently, $(\vartheta, r)$) as known. To be fair, when we compare CASE with other methods, we also assume $(\epsilon_p, \tau_p)$ as known when we set the tuning parameters for the latter.

In light of this, we set $m = 3$ when $\vartheta < 0.3$, $m = 2$ when $0.3 \leq \vartheta < 0.5$, and $m = 1$ otherwise. In this setting, any $\delta \in (0, 1)$ gives the same graph $\mathcal{G}^*$, so we take $\delta = 0.5$. Additionally, we set $u^{pe} = \sqrt{2\log(1/\epsilon_p)}$ and $v^{pe} = \tau_p$. The choice of $\ell^{pe}$ is heuristic and depends on how small $p^{-\vartheta}$ is; in our numerical studies, $\ell^{pe}$ ranges from 10 to 35 in the case $p = 5000$ for different $\vartheta$, and it ranges from 20 to 200 in the case $p = 10^6$. Last, we take the patching method as described in (1.42) and then apply the $PE$-step.

*Experiment 1.1(a).* In this experiment, we misspecify $(\epsilon_p, \tau_p)$, say, as $(\tilde{\epsilon}_p, \tilde{\tau}_p)$, and set $u^{pe} = \sqrt{2\log(1/\tilde{\epsilon}_p)}$ and $v^{pe} = \tilde{\tau}_p$, and investigate how the misspecification affects the performance of CASE. Fix $(p, \vartheta, \tau_p, a) = (5000, 0.60, 5, 1)$, so that $(\epsilon_p, \tau_p) = (0.006, 5)$. We misspecify $(\epsilon_p, \tau_p)$ by a small amount where we let $\tilde{\tau}_p$ vary in $\{4, 4.5, \cdots, 6\}$, and let $\tilde{\epsilon}_p$ vary in $\{0.005, 0.0055, \cdots, 0.007\}$. Table 1 reports the average Hamming errors of 50 independent repetitions. The results suggest that CASE is reasonably insensitive to the misspecification: the performance of CASE where $(\epsilon_p, \tau_p)$ are misspecified is close to the case where $(\epsilon_p, \tau_p)$ are assumed as known.

For comparison, we also investigate the performance of SaRa (see [24]), which is defined as

$$\hat{\beta}_i^{SaRa} = W_i 1\{|W_i| > \lambda\}, \qquad \text{where} \qquad W_i = \frac{1}{h}(\sum_{j=i+1}^{i+h} Y_j - \sum_{j=i-h+1}^{i} Y_j).$$

SaRa uses two tuning parameters $(h, \lambda)$ which we set ideally assuming $(\epsilon_p, \tau_p)$ as known: for all $(h, \lambda)$ satisfying $h \leq \lfloor 1/\epsilon_p \rfloor$ and $\lambda \leq \tau_p$, we choose (by exhaustive numerical search) the pair that yields the smallest Hamming error. In this setting, the average Hamming error of SaRa is 9.02 (compare Table 1). We see that CASE consistently outperforms SaRa, even when CASE uses the misspecified $(\epsilon_p, \tau_p)$ to determine the tuning parameters $(u^{pe}, v^{pe})$,

TABLE 1
*Hamming errors in Experiment 1.1(a). $p = 5000$, $\vartheta = 0.60$ and $\tau_p = 5$. The expected number of signals is $p^{1-\vartheta} = 30$.*

| $\tilde{\epsilon}_p$ | $\tilde{\tau}_p$ | | | | |
|---|---|---|---|---|---|
| | 4 | 4.5 | 5 | 5.5 | 6 |
| 0.0050 | 5.50 | 5.26 | 5.04 | 5.08 | 5.22 |
| 0.0055 | 5.10 | 5.04 | 4.84 | 4.82 | 5.12 |
| 0.0060 | 5.02 | 4.82 | 4.78 | 4.74 | 4.98 |
| 0.0065 | 5.06 | 4.86 | 4.78 | 4.76 | 4.98 |
| 0.0070 | 5.26 | 4.96 | 4.84 | 4.84 | 5.00 |

and SaRa uses the true values of $(\epsilon_p, \tau_p)$ to determine the tuning parameters $(\lambda, h)$.

*Experiment 1.1(b).* In this experiment, we investigate the performance of CASE when $(\epsilon_p, \tau_p)$ are unknown but can be estimated. We propose the following approach to estimate $(\epsilon_p, \tau_p)$:

$$\hat{\epsilon}_p = \frac{1}{p}\sum_{i=1}^{p} 1\{|W_i| > \lambda\}, \qquad \hat{\tau}_p = \frac{1}{p\hat{\epsilon}_p}\sum_{i=1}^{p} |W_i| 1\{|W_i| > \lambda\},$$

where $W_i = \frac{1}{h}(\sum_{j=i+1}^{i+h} Y_j - \sum_{j=i-h+1}^{i} Y_j)$, and $(\lambda, h)$ are tuning parameters. Our numerical studies find that the approach works satisfactorily, especially when $\tau_p$ is moderately large and $\epsilon_p$ is moderately small.

Fix $(p, a, \lambda, h) = (5000, 1, 4.5, 5)$. We investigate different settings with $\vartheta \in \{0.60, 0.45\}$ and $\tau_p \in \{4, \cdots, 9\}$. We compare the performance of CASE where $(u^{pe}, v^{pe})$ are computed based on $(\hat{\epsilon}_p, \hat{\tau}_p)$, CASE where $(u^{pe}, v^{pe})$ are computed based on $(\epsilon_p, \tau_p)$, and SaRa. Figure 4 summarizes the results based on 50 independent repetitions. The results suggest that two versions of the CASE have similar performance, which is substantially better than that of SaRa.

*Experiment 1.2.* In this experiment, we compare CASE with the naive hard thresholding (nHT) introduced in Section 1.11. The tuning parameters of CASE are set in a way assuming $(\tau_p, \epsilon_p)$ as known. The threshold of nHT is set ideally as $(r + 2\vartheta)^2/(2r) \cdot \log(p)$ (where we also assume that $(\epsilon_p, \tau_p)$ as known). Fix $p = 10^6$ and $a = 1$. Let $\vartheta$ range in $\{0.35, 0.5, 0.75\}$, and $\tau_p$ range in $\{5, \cdots, 13\}$. Figure 5 summaries the average Hamming errors of 50 independent repetitions. The results suggest that CASE outperforms the naive hard thresholding.

*Experiment 1.3.* In this experiment, we compare the performance of three procedures, CASE, SaRa and the lasso, with a few representative pairs of $(\vartheta, \tau_p)$. Note here that the lasso estimate, $\hat{\beta}^{lasso}$, is the minimizer of the

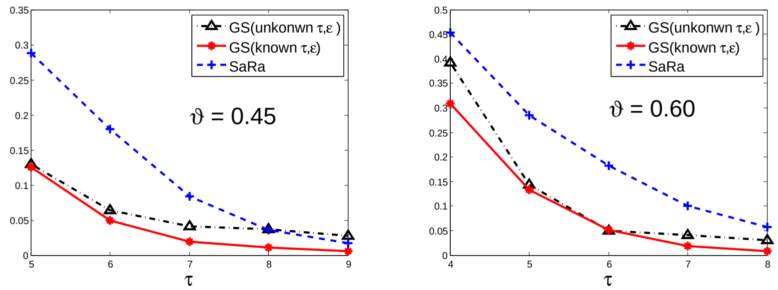



FIG 4. *Hamming errors in Experiment 1.1(b) ($p = 5000$). The $x$ axis is $\tau_p$, and the $y$ axis is the ratio between Hamming error and $p^{1-\vartheta}$.*

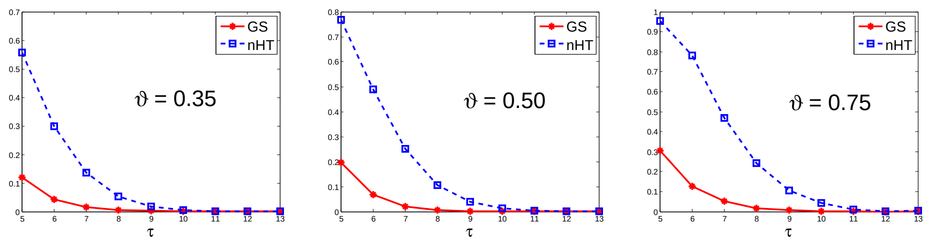



FIG 5. *Hamming errors in Experiment 1.2 ($p = 10^6$). The $x$ axis is $\tau_p$, and the $y$ axis is the ratio between Hamming error and $p^{1-\vartheta}$.*

following functional

$$\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \lambda\|\beta\|_1,$$

where $\lambda > 0$ is a tuning parameter. We use the the *glmnet* package [13] in the simulations.

Fix $p = 5000$ and $a = 1$. We let $\vartheta$ range in $\{0.3, 0.45, 0.65\}$ and $\tau_p$ range in $\{3, 4, \cdots, 10\}$. The tuning parameters of CASE and SaRa are set ideally as in Experiment 1.1(a), assuming $(\epsilon_p, \tau_p)$ as known. The lasso tuning parameter $\lambda$ is also set ideally (we calculate the whole solution path and choose the one with the smallest Hamming error). Table 2 displays the results based on 50 independent repetitions, which suggests that CASE outperforms the other two methods in most cases.

In particular, the lasso behaves unsatisfactorily, due to the strong dependence among the design variables. Similar conclusion can be drawn in most of the examples considered in the section, but to save space, we only report that of the lasso here.

*Experiment 1.4.* In this experiment, we let $a > 1$ so the signals may

TABLE 2
*Hamming errors in Experiment 1.3 ($p = 5000$).*

| $\vartheta$ | $s_p$ | | $\tau_p$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0.30 | 388.4 | CASE | 212.8 | 106.9 | 52.4 | 25.1 | 14.8 | 9.90 | 7.64 | 6.66 |
| | | lasso | 375.2 | 373.3 | 374.9 | 371.9 | 372.6 | 378.1 | 369.9 | 374.0 |
| | | SaRa | 245.3 | 175.1 | 106.4 | 50.4 | 21.2 | 6.12 | 2.98 | 1.38 |
| 0.45 | 108.3 | CASE | 56.6 | 28.9 | 11.6 | 4.70 | 1.68 | 0.82 | 0.72 | 0.64 |
| | | lasso | 105.4 | 106.1 | 105.8 | 102.6 | 103.1 | 106.2 | 103.7 | 105.1 |
| | | SaRa | 76.0 | 48.3 | 31.1 | 18.3 | 9.16 | 3.84 | 1.76 | 1.06 |
| 0.65 | 19.7 | CASE | 11.8 | 5.94 | 2.36 | 0.96 | 0.38 | 0.18 | 0.12 | 0.14 |
| | | lasso | 19.7 | 18.3 | 18.8 | 19.2 | 19.1 | 20.1 | 20.1 | 19.4 |
| | | SaRa | 14.5 | 8.50 | 5.42 | 3.94 | 2.16 | 1.42 | 1.06 | 1.00 |

TABLE 3
*Hamming errors in Experiment 1.4. $p = 5000$, $\vartheta = 0.5$, $p^{1-\vartheta} = 70.7$ and $\tau_p = 4.5$.*

| | | $a$ | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 1.5 | 2 | 2.5 | 3 |
| half-half | CASE | 14.26 | 6.32 | 5.50 | 4.78 | 4.56 |
| | SaRa | 24.98 | 18.96 | 16.56 | 14.00 | 12.50 |
| all-positive | CASE | 13.44 | 6.18 | 4.90 | 5.38 | 4.14 |
| | SaRa | 24.26 | 18.58 | 16.80 | 13.66 | 12.12 |

have different strengths. Fix $(p, \vartheta, \tau_p) = (5000, 0.50, 4.5)$, and let $a$ range in $\{1, 1.5, \cdots, 4\}$. We investigate a case where the signals have the "half-positive-half-negative" sign pattern, i.e., $\beta_j \overset{iid}{\sim} (1 - \epsilon_p)\nu_0 + \frac{\epsilon_p}{2} U(\tau_p, a\tau_p) + \frac{\epsilon_p}{2} U(-a\tau_p, -\tau_p)$, and a case where the the signals have the "all-positive" sign pattern, i.e., $\beta_j \overset{iid}{\sim} (1 - \epsilon_p)\nu_0 + \epsilon_p U(\tau_p, a\tau_p)$. We compare CASE with SaRa for different values of $a$ and sign-patterns. The results of 50 independent repetitions are reported in Table 3, which suggest that CASE uniformly outperforms SaRa for various values of $a$ and the two sign patterns.

3.2. *Long-memory time series model.* In this section, we consider the long-memory time seris model with a specific $f$ as in (1.39) and (1.40). Fix $(p, \phi, \vartheta, \tau_p, a)$, where $\phi$ is the long-memory parameter. We first use $f$ to compute $G$ and let $X = G^{1/2}$. We then generate the vector $\beta$ by $\beta_j \overset{iid}{\sim} (1 - \epsilon_p)\nu_0 + \frac{\epsilon_p}{2} U(\tau_p, a\tau_p) + \frac{\epsilon_p}{2} U(-a\tau_p, -\tau_p)$. Finally, we generate $Y \sim N(X\beta, I_p)$.

CASE uses tuning parameters $(m, \delta, \ell^{ps}, \mathcal{Q}, \ell^{pe}, u^{pe}, v^{pe})$. In experiments below, we choose them as follows: $m = 2$, $\delta = 0.35$, $u^{pe} = \sqrt{2\vartheta \log(p)}$ and $v^{pe} = \sqrt{2r \log(p)}$. We take $t(\hat{F}, \hat{N}) = q^*(\hat{F}, \hat{N}) \log(p)$, where $q^*(\hat{F}, \hat{N})$ is

TABLE 4
*Hamming errors in Experiment 2.1 ($\phi = 0.35$, $p = 5000$).*

| $\vartheta$ | $s_p$ | | $\tau_p$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 0.35 | 253.7 | CASE | 130.6 | 66.9 | 29.4 | 11.1 | 4.20 | 2.26 | 1.36 | 1.42 |
| | | lasso | 144.2 | 96.0 | 62.2 | 38.1 | 24.0 | 17.4 | 12.0 | 9.1 |
| 0.50 | 70.7 | CASE | 45.6 | 24.8 | 11.3 | 3.76 | 1.52 | 0.68 | 0.52 | 0.72 |
| | | lasso | 42.0 | 25.8 | 14.1 | 7.54 | 4.44 | 2.00 | 1.38 | 1.12 |
| 0.65 | 19.7 | CASE | 12.2 | 6.56 | 2.48 | 0.76 | 0.38 | 0.22 | 0.06 | 0.12 |
| | | lasso | 11.4 | 6.24 | 2.74 | 1.14 | 0.36 | 0.18 | 0.08 | 0.00 |

defined in (1.31), and $5 \leq \ell^{ps} \leq 10$, depending on how large $\vartheta$ is (small $\vartheta$ corresponds to large $\ell^{ps}$). $\ell^{pe}$ is chosen in this way: for a certain range of integers, run the CASE for each and choose the largest integer such that each component of $\mathcal{U}_p^*$ (as a subgraph $\mathcal{G}^+$) has a size $\leq 10$. In general, larger $\ell^{pe}$ has better performance, but may result in longer computation time.

*Experiment 2.1.* Fix $p = 5000$, $\phi = 0.35$ and $a = 1$. Let $\vartheta$ range in $\{0.35, 0.5, 0.65\}$, and $\tau_p$ range in $\{4, \cdots, 11\}$. We compare the performance of CASE with that of the lasso. The tuning parameters of CASE are set as above. The tuning parameters of the lasso are the oracle ones as in Experiment 1.3. The results based on 50 independent repetitions are summarized in Table 4. We see that CASE uniformly outperforms the lasso when $\vartheta = 0.35, 0.5$. When $\vartheta = 0.65$, the performances of the two methods are similar.

*Experiment 2.2.* In this experiment, we force the signals to appear in adjacent pairs or triplets. Fix $p = 5000$, $\phi = 0.35$, $\vartheta = 0.75$ and let $\tau_p$ range in $\{5, \cdots, 10\}$. We use '$+-$' to denote the signal pattern 'pairs of opposite signs', '$++$' 'pairs of the same sign'. Other signal patterns are denoted similarly. To generate $\beta$ corresponding to '$+-$', we first generate a $(p/2) \times 1$ vector $\theta$ by $\theta_j \stackrel{iid}{\sim} (1 - \epsilon_p)\nu_0 + \frac{\epsilon_p}{2} U(\tau_p, a\tau_p) + \frac{\epsilon_p}{2} U(-a\tau_p, -\tau_p)$, then let $\beta_{2j-1} = \theta_j$ and $\beta_{2j} = \theta_j$. Similarly for other signal patterns. Figure 6 displays the results of 50 independent repetitions. We see that in the four patterns '$+-$', '$++-$', '$+-+$' and '$+--$', CASE uniformly outperforms the lasso when $\tau_p \geq 6$.

**4. Discussion.** Variable selection when the Gram matrix $G$ is non-sparse is a challenging problem. We approach this problem by first sparsifying $G$ with a finite order linear filter, and then constructing a sparse graph GOSD. The key insight is that, in the post-filtering data, the true signals live in many small-size components that are disconnected in GOSD, but we do not know where. We propose CASE as a new approach to variable

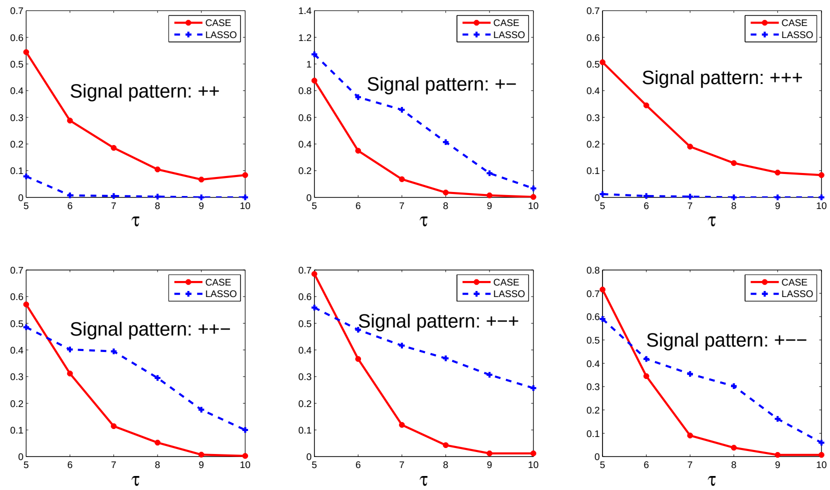



FIG 6. *Hamming errors in Experiment 2.2.*

selection. This is a two-stage Screen and Clean method, where we first use a covariance-assisted multivariate screening to identify candidates for such small-size components, and then re-examine each candidate with penalized least squares. In both stages, to overcome the problem of information leakage, we employ a delicate patching technique.

We develop an asymptotic framework focusing on the regime where the signals are rare and weak so that successful variable selection is challenging but is still possible. We show that CASE achieves the optimal rate of convergence in Hamming distance across a wide class of situations where $G$ is non-sparse but sparsifiable. Such optimality cannot be achieved by many popular methods, including but not limited to the lasso, SCAD, and Dantzig selector. When $G$ is non-sparse, these methods are not expected to behave well even when the signals are strong. We have successfully applied CASE to two different applications: the change-point problem and the long-memory times series.

Compared to the well-known method of marginal screening [10, 32], CASE employs a covariance-assisted multivariate screening procedure, so that it is theoretically more effective than marginal screening, with only a moderate increase in the computational complexity. CASE is closely related to the graphical lasso [12, 22], which also attempts to exploit the graph structure. However, the setting considered here is very different from that in [12, 22] and our emphasis on optimality is also very different.

The paper is closely related to the recent work [20] (see also [19]), but is

different in important ways. The work in [20] is motivated by recent literature of Compressive Sensing and Genetic Regulatory Network, and is largely focused on the case where the Gram matrix $G$ is sparse in an unstructured fashion. The current work is motivated by the recent interest on DNA-copy number variation and long-memory time series, and is focused on the case where there are strong dependence between different design variables so $G$ is usually non-sparse and some times ill-posed. To deal with the strong dependence, we have to use a finite-order linear filter and delicate patching techniques. Additionally, the current paper also studies applications to the long-memory time series and change-point problem which have not been considered in [20]. Especially, the studies on the change-point problem encompasses very different and very delicate analysis on both the derivation of the lower bound and upper bound which we have not seen before in the literature. For these reasons, the two papers have very different scopes and techniques, and the results in one paper cannot be deduced from those in the other.

The main results in this paper can be extended to much broader settings. For example, we have used a Rare and Weak signal model where the signals are randomly generated from a two-component mixture. The main results continue to hold if we choose to use a much more relaxed model, as long as the signals live in small-size isolated islands in the post-filtering data.

In this paper, we have focused on the change-point model and the long-memory time series model, where the post-filtering matrices have polynomial off-diagonal decay and are sparse in a structured fashion. CASE can be extended to more general settings, where the sparsity of the post-filtering matrices are unstructured, provided that we modify the patching technique accordingly: the patching set can be constructed by including nodes which are connected to the original set through a short-length path in the GOSD $\mathcal{G}^*$.

Another extension is that the Gram matrix can be sparsified by an operator $D$, but $D$ is not necessary linear filtering. To apply CASE to this setting, we need to design specific patching technique. For example, when $D^{-1}$ is sparse, for a given $\mathcal{I}$, we can construct $\mathcal{I}^{pe} = \{j : |D^{-1}(i,j)| > \delta_1, \text{ for some } i \in \mathcal{I}\}$, where $\delta_1$ is a chosen threshold.

The paper is closely related to recent literature on DNA copy number variation and financial data analysis, but is different in focus and scope. It is of interest to further investigate such connections. To save space, we leave explorations along this line to the future.

**5. Proofs.** This section is organized as follows. In Section 5.1, we state and prove three preliminary lemmas, which are useful for this section. In Sections 5.2-5.12, we give the proofs of all the main theorems and lemmas stated in the preceding sections.

5.1. *Preliminary lemmas.* We introduce Lemmas 5.1-5.3, where Lemmas 5.1-5.2 are proved below, and Lemma 5.3 is proved in [20, Lemma 1.4].

Recall that $B = DG$ and $\mathcal{G}^*$ is the GOSD in Definition 1.3 with $\delta = 1/\log(p)$. Introduce the matrix $B^{**}$ by

$$B^{**}(i,j) = B(i,j) \cdot 1\big\{j \in \mathcal{E}(\{i\})\big\}, \qquad 1 \le i,j \le p,$$

where for any set $V \subset \{1, \cdots, p\}$,

$$\mathcal{E}(V) = \big\{k: \text{ there is an edge between } k \text{ and } k' \text{ in } \mathcal{G}^* \text{ for some } k' \in V\big\}.$$

Recall that $\mathcal{M}_p(\alpha, A_0)$ is the class of matrices defined in (1.7).

LEMMA 5.1. *When $B \in \mathcal{M}_p(\alpha, A_0)$, $\mathcal{G}^*$ is $K_p$-sparse for $K_p \le C[\log(p)]^{1/\alpha}$, and $\|B - B^{**}\|_\infty \le C[\log(p)]^{-(1-1/\alpha)}$.*

PROOF. Consider the first claim. Since $B \in \mathcal{M}_p(\alpha, A_0)$ and $H(i,j) = \sum_{k=0}^h \eta_k B(i, j+k)$, there exists a constant $A_0' > 0$ such that $H \in \mathcal{M}_p(\alpha, A_0')$. Let $K_p$ be the smallest integer satisfying

$$K_p \ge 2[\max(A_0, A_0') \log(p)]^{1/\alpha},$$

where it is seen that $K_p \le C(\log(p))^{1/\alpha}$. At the same time, for any $i, j$ such that $|i-j|+1 > K_p/2$, we have $|B(i,j)| < \delta$, $|B(j,i)| < \delta$ and $|H(i,j)| < \delta$. By definition, there is no edge between nodes $i$ and $j$ in $\mathcal{G}^*$. This proves that $\mathcal{G}^*$ is $K_p$-sparse, and the claim follows.

Consider the second claim. When $|B(i,j)| > \delta$, there is an edge between nodes $i$ and $j$ in $\mathcal{G}^*$, and it follows that $(B - B^{**})(i,j) = 0$. Therefore, for any $1 \le i \le p$,

$$\begin{aligned} \sum_{j=1}^p |(B - B^{**})(i,j)| &\le \sum_{j:|j-i|+1 > K_p/2} |B(i,j)| + \sum_{j:|j-i|+1 \le K_p/2, |B(i,j)| \le \delta} |B(i,j)| \\ &\equiv I + II, \end{aligned}$$

where $I \le 2A_0 \sum_{k+1 > K_p/2} k^{-\alpha} \le CK_p^{1-\alpha}$ and $II \le K_p \delta = CK_p^{1-\alpha}$. Recalling $K_p \le C[\log(p)]^{\alpha}$, $\|B - B^{**}\|_\infty \le CK_p^{1-\alpha} \le C[\log(p)]^{-(1-1/\alpha)}$, and the claim follows. □

Next, recall that $\mathcal{G}^+$ is an expanded graph of $\mathcal{G}^*$, given in Definition 1.7, and $\mathcal{I} \lhd \mathcal{G}$ denotes that $\mathcal{I}$ is a component of $\mathcal{G}$, as in Definition 1.8.

LEMMA 5.2. *When $\mathcal{G}^*$ is $K$-sparse, $\mathcal{G}^+$ is $K(2\ell^{pe}+1)^2$-sparse. In addition, for any set $V \subset \{1, \cdots, p\}$, let $\mathcal{G}^+_V$ be the subgraph of $\mathcal{G}^+$ formed by nodes in $V$. Then for any $\mathcal{I} \lhd \mathcal{G}^+_V$, $(V \backslash \mathcal{I}) \cap \mathcal{E}(\mathcal{I}^{pe}) = \emptyset$.*

PROOF. Consider the first claim. It suffices to show that for any fixed $1 \le i \le p$, there are at most $K(2\ell^{pe}+1)^2$ different nodes $j$ such that there is an edge between $i$ and $j$ in $\mathcal{G}^+$. Towards this end, note that $\{i\}^{pe}$ contains no more than $(2\ell^{pe}+1)$ nodes. Since $\mathcal{G}^*$ is $K$-sparse, for each $k \in \{i\}^{pe}$, there are no more than $K$ nodes $k'$ such that there is an edge between $k$ and $k'$ in $\mathcal{G}^*$. Again, for each such $k'$, there are no more than $(2\ell^{pe}+1)$ nodes $j$ such that $k' \in \{j\}^{pe}$. Combining these gives the claim.

Consider the second claim. Fix $V$ and $\mathcal{I} \lhd \mathcal{G}^+_V$. Since $\mathcal{I}$ is a component, for any $i \in \mathcal{I}$ and $j \in V \backslash \mathcal{I}$, there is no edge between $i$ and $j$ in $\mathcal{G}^+_V$. By definition, this implies $\{j\}^{pe} \cap \mathcal{E}(\{i\}^{pe}) = \emptyset$, and especially $j \notin \mathcal{E}(\{i\}^{pe})$. Since this holds for all such $i$ and $j$, using that $\mathcal{E}(\mathcal{I}^{pe}) = \cup_{i \in \mathcal{I}} \mathcal{E}(\{i\}^{pe})$, we have $(V \backslash \mathcal{I}) \cap \mathcal{E}(\mathcal{I}^{pe}) = \emptyset$, and the claim follows. $\square$

Finally, recall the definition of $\rho_j^*(\vartheta, r, a, G)$ in (1.27) and that of $\psi(F, N)$ in (1.33).

LEMMA 5.3. *When $a > a_g^*(G)$, $\rho_j^*(\vartheta, r, a, G)$ does not depend on $a$ and $\rho_j^*(\vartheta, r, a, G) \equiv \rho_j^*(\vartheta, r, G) = \min_{(F,N): j \in F, F \cap N = \emptyset, F \neq \emptyset} \psi(F, N)$.*

5.2. *Proof of Lemma 1.2.* For preparation, note that the Fisher Information Matrix associated with model (1.13) is

$$Q \equiv (B^{\mathcal{I}^+, \mathcal{I}})'(H^{\mathcal{I}^+, \mathcal{I}^+})^{-1}(B^{\mathcal{I}^+, \mathcal{I}}).$$

Write $D_1 = D^{\mathcal{I}^+, \mathcal{J}^+}$ and $G_1 = G^{\mathcal{J}^+, \mathcal{J}^+}$ for short. It follows that $B^{\mathcal{I}^+, \mathcal{J}^+} = D_1 G_1$ and $H^{\mathcal{I}^+, \mathcal{I}^+} = D_1 G_1 D_1'$. Let $\mathcal{F}$ be the mapping from $\mathcal{J}^+$ to $\{1, \cdots, |\mathcal{J}^+|\}$ that maps each $j \in \mathcal{J}^+$ to its order in $\mathcal{J}^+$, and let $\mathcal{I}_1 = \mathcal{F}(\mathcal{I})$. By these notations, we can write

$$Q = Q_1^{\mathcal{I}_1, \mathcal{I}_1}, \qquad \text{where} \qquad Q_1 \equiv G_1 D_1'(D_1 G_1 D_1')^{-1} D_1 G_1. \tag{5.67}$$

Comparing (5.67) with the desired claim, it suffices to show

$$Q_1 = G_1 - U(U' G_1^{-1} U)^{-1} U'. \tag{5.68}$$

Let $R = D_1 G_1^{1/2}$ and $P_R = R'(RR')^{-1}R$. It is seen that

$$Q_1 = G_1^{1/2} P_R G_1^{1/2} = G_1 - G_1^{1/2}(I - P_R)G_1^{1/2}. \tag{5.69}$$

Now, we study the matrix $I - P_R$. Let $k = |\mathcal{J}^+|$, and denote $\mathcal{S}(R)$ the row space of $R$ and $\mathcal{N}(R)$ the orthogonal complement of $\mathcal{S}(R)$ in $\mathbb{R}^k$. By construction, $P_R$ is the orthogonal projection matrix from $\mathbb{R}^k$ to $\mathcal{S}(R)$. Hence, $I - P_R$ is the orthogonal projection matrix from $\mathbb{R}^k$ to $\mathcal{N}(R)$. By definition, $\mathcal{N}(R) = \{\eta \in \mathbb{R}^k : R\eta = 0\}$. Recall that $R = D_1 G_1^{1/2}$. Therefore, $R\eta = 0$ if and only if there exists $\xi \in \mathbb{R}^k$ such that $\eta = G_1^{-1/2}\xi$ and $D_1\xi = 0$. At the same time, $Null(\mathcal{I}^+, \mathcal{J}^+) = \{\xi \in \mathbb{R}^k : D_1\xi = 0\}$. Combining these, we have

$$\mathcal{N}(R) = \big\{G_1^{-1/2}\xi :\ \xi \in Null(\mathcal{I}^+, \mathcal{J}^+)\big\}. \tag{5.70}$$

Introduce a new matrix $V = G_1^{-1/2}U$. Since the columns of $U$ form an orthonormal basis of $Null(\mathcal{I}^+, \mathcal{J}^+)$, it follows from (5.70) that the columns of $V$ form a basis (but not necessarily an orthonormal basis) of $\mathcal{N}(R)$. Consequently,

$$I - P_R = V(V'V)^{-1}V' = G_1^{-1/2}U(U'G_1^{-1}U)^{-1}U'G_1^{-1/2}. \tag{5.71}$$

Plugging (5.71) into (5.69) gives (5.68). $\square$

5.3. *Proof of Lemma 1.4.* Write $\rho_j^* = \rho_j^*(\vartheta, r, G)$ for short. It suffices to show for any $\log(p) \leq j \leq p - \log(p)$, there exists $(V_0, V_1)$ such that

$$\rho(V_0, V_1) \leq \rho_j^* + o(1), \qquad j \in (V_0 \cup V_1) \subset \{j + i : -\log(p) \leq i \leq \log(p)\}. \tag{5.72}$$

In fact, once (5.72) is proved, then $d_p(\mathcal{G}^\diamond) \leq 2\log(p) + 1$, and the claim follows directly.

We now construct $(V_0, V_1)$ to satisfy (5.72) for any $j$ such that $\log(p) \leq j \leq p - \log(p)$. The key is to construct a sequence of set pairs $(V_0^{(t)}, V_1^{(t)})$ recursively as follows. Let $V_0^{(1)} = V_{0j}^*$ and $V_1^{(1)} = V_{1j}^*$, where $(V_{0j}^*, V_{1j}^*)$ are as defined in Section 1.8. For any integer $t \geq 1$, we update $(V_0^{(t)}, V_1^{(t)})$ as follows. If all inter-distance between the nodes in $V_0^{(t)} \cup V_1^{(t)}$ (assuming all nodes are sorted ascendingly) does not exceed $\log(p)/g$, then the process terminates. Otherwise, there are a pair of adjacent nodes $i_1$ and $i_2$ in $(V_0^{(t)} \cup V_1^{(t)})$ (again, assuming the nodes are sorted ascendingly) such that $i_2 > i_1 + \log(p)/g$. In our construction, it is not hard to see that $j \in V_0^{(t)} \cup V_1^{(t)}$. Therefore, we have either the case of $j \leq i_1$ or the case of $j \geq i_2$. In the first case, we let

$$N^{(t+1)} = N^{(t)} \cap \{i : i \leq i_1\}, \qquad F^{(t+1)} = F^{(t)} \cap \{i : i \leq i_1\},$$

and in the second case, we let

$$N^{(t+1)} = N^{(t)} \cap \{i : i \geq i_2\}, \qquad F^{(t+1)} = F^{(t)} \cap \{i : i \geq i_2\},$$

where $N^{(t)} = V_0^{(t)} \cap V_1^{(t)}$ and $F^{(t)} = (V_0^{(t)} \cup V_1^{(t)}) \setminus N^{(t)}$. We then update by defining

$$V_0^{(t+1)} = N^{(t+1)} \cup F', \qquad V_1^{(t+1)} = N^{(t+1)} \cup F''$$

where $(F', F'')$ are constructed as follows: Write $F^{(t)} = \{j_1, j_2, \cdots, j_k\}$ where $j_1 < j_2 < \ldots < j_k$ and $k = |F^{(t)}|$. When $k$ is even, let $F' = \{j_1, \cdots, j_{k/2}\}$ and $F'' = F^{(t)} \backslash F'$; otherwise, let $F' = \{j_1, \cdots, j_{(k-1)/2}\}$ and $F'' = F^{(t)} \backslash F'$.

Now, first, by the construction, $|F^{(t)} \cup N^{(t)}|$ is strictly decreasing in $t$. Second, by [20, Lemma 1.2], $|F^{(1)} \cup N^{(1)}| \leq |V_{0j}^* \cup V_{1j}^*| \leq g$. As a result, the recursive process above terminates in finite rounds. Let $T$ be the number of rounds when the process terminates, we construct $(V_0, V_1)$ by

$$V_0 = V_0^{(T)}, \qquad V_1 = V_1^{(T)}. \tag{5.73}$$

Next, we justify $(V_0, V_1)$ constructed in (5.73) satisfies (5.72). First, it is easy to see that $j \in V_0 \cup V_1$ and $|V_0 \cup V_1| \leq g$. Second, all pairs of adjacent nodes in $V_0 \cup V_1$ have an inter-distance $\leq \log(p)/g$ (assuming all nodes are sorted), so $(V_0 \cup V_1) \subset \{j - \log(p), \cdots, j + \log(p)\}$. As a result, all remains to show is

$$\rho(V_0, V_1) \leq \rho_j^* + o(1). \tag{5.74}$$

By similar argument as in [20, Lemma 1.4] and definitions (i.e. (1.33) and [20, (1.23)]), if $a > a_g^*(G)$, then for any $(V_0', V_1')$ such that $|V_0' \cup V_1'| \leq g$, we have $\rho(V_0', V_1') \geq \psi(F', N')$, where $N' = V_0' \cap V_1'$ and $F' = (V_0' \cup V_1') \backslash N'$. Moreover, the equality holds when $|V_0'| = |V_1'|$ in the case $|F'|$ is even, and $|V_0'| - |V_1'| = \pm 1$ in the case $|F'|$ is odd. Combining these with definitions,

$$\rho(V_0, V_1) = \psi(F^{(T)}, N^{(T)}), \quad \rho_j^* \equiv \rho(V_{0j}^*, V_{1j}^*) = \rho(V_0^{(1)}, V_1^{(1)}) \geq \psi(F^{(1)}, N^{(1)}).$$

Recall that $T$ is a finite number. So to show (5.74), it suffices to show for each $1 \leq t \leq T - 1$,

$$\psi(F^{(t+1)}, N^{(t+1)}) \leq \psi(F^{(t)}, N^{(t)}) + o(1). \tag{5.75}$$

Fixing $1 \leq t \leq T-1$, write for short $F = F^{(t)}$, $N = N^{(t)}$, $N_1 = N^{(t+1)}$ and $F_1 = F^{(t+1)}$. Let $\mathcal{I} = F \cup N$ and $\mathcal{I}_1 = F_1 \cup N_1$. With these notations, (5.75) reduces to

$$\psi(F_1, N_1) \leq \psi(F, N) + o(1). \tag{5.76}$$

By the way $\psi$ is defined (i.e., (1.33)), it is sufficient to show

$$\omega(F_1, N_1) \leq \omega(F, N) + o(1). \tag{5.77}$$

In fact, once (5.77) is proved, (5.75) follows by noting that $|F_1| + 2|N_1| \leq |F| + 2|N| - 1$.

We now show (5.77). Letting $\Omega = \mathrm{diag}(G^{\mathcal{I}_1,\mathcal{I}_1}, G^{\mathcal{I}\setminus\mathcal{I}_1,\mathcal{I}\setminus\mathcal{I}_1})$, we write

$$\begin{aligned}
\omega(F,N) &= \min_{\theta\in\mathbb{R}^{|\mathcal{I}|}:|\theta_i|\geq 1,\forall i\in F} \theta' G^{\mathcal{I},\mathcal{I}}\theta \\
&\geq \min_{\theta\in\mathbb{R}^{|\mathcal{I}|}:|\theta_i|\geq 1,\forall i\in F} \theta'\Omega\theta - \max_{\theta\in\mathbb{R}^{|\mathcal{I}|}:|\theta_i|\leq 2a,\forall i} |\theta'(G^{\mathcal{I},\mathcal{I}} - \Omega)\theta| \\
&\geq \min_{\theta\in\mathbb{R}^{|\mathcal{I}_1|}:|\theta_i|\geq 1,\forall i\in F_1} \theta' G^{\mathcal{I}_1,\mathcal{I}_1}\theta - \max_{\theta\in\mathbb{R}^{|\mathcal{I}|}:|\theta_i|\leq 2a,\forall i} |\theta'(G^{\mathcal{I},\mathcal{I}} - \Omega)\theta| \qquad (5.78)\\
&= \omega(F_1,N_1) - \max_{\theta\in\mathbb{R}^{|\mathcal{I}|}:|\theta_i|\leq 2a,\forall i} |\theta'(G^{\mathcal{I},\mathcal{I}} - \Omega)\theta|,
\end{aligned}$$

where in the first and last equalities we use equivalent forms of $\omega(F, N)$, in the second inequality we use the fact that the constraints $|\theta_i| \geq 1$ can be replaced by $1 \leq |\theta_i| \leq a$ for any $a > a_g^*$ and the triangular inequality, and in the third inequality we use the definition of $\Omega$.

Finally, note that for any $k \in \mathcal{I}_1$ and $k' \in \mathcal{I}\setminus\mathcal{I}_1$, $|k - k'| > \log(p)/g$ holds. In addition, $G$ has polynomial off-diagonal decays with rate $\gamma > 0$. Together we find that $\|G^{\mathcal{I},\mathcal{I}} - \Omega\| \leq C(\log(p)/g)^{-\gamma} = o(1)$. As a result, $\max_{\theta\in\mathbb{R}^{|\mathcal{I}|}:|\theta_i|\leq 2a,\forall i} |\theta'(G^{\mathcal{I},\mathcal{I}} - \Omega)\theta| \leq Ca^2 \cdot \|G^{\mathcal{I},\mathcal{I}} - \Omega\| \cdot |\mathcal{I}| = o(1)$. Inserting this into (5.78) gives (5.77). ☐

5.4. *Proof of Theorem 1.3.* First, we define $\rho_{lts}^*(\vartheta, r; f)$ as follows. For any spectral density function $f$, let $G^\infty = G^\infty(f)$ be the (infinitely dimensional) Toeplitz matrix generated by $f$: $G^\infty(i,j) = \hat{f}(|i-j|)$ for any $i, j \in \mathbb{Z}$, where $\hat{f}(k)$ is the $k$-th Fourier coefficient of $f$. In the definition of $\rho(V_0, V_1)$ in (1.25)-(1.26), replace $G$ by $G^\infty$ and call the new term $\rho^\infty(V_0, V_1)$. For any fixed $j$, let

$$\rho_{j,lts}^*(\vartheta, r; f) = \min_{(V_0,V_1):j\in V_0\cup V_1} \rho^\infty(V_0, V_1), \tag{5.79}$$

where $V_0, V_1$ are subsets of $\mathbb{Z}$. Due to the definition of Toeplitz matrices, $\rho_{j,lts}^*(\vartheta, r; f)$ does not depend on $j$, so we write it as $\rho_{lts}^*(\vartheta, r; f)$ for short. By (5.72), it is seen that

$$\rho_j^*(\vartheta, r, G) = \rho_{lts}^*(\vartheta, r; f) + o(1), \quad \text{for any } \log(p) \leq j \leq p - \log(p). \tag{5.80}$$

Now, to show the claim, it is sufficient to check the main conditions of Theorem 1.2. In detail, it suffices to check that

(a) $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$ with $\gamma = 1 - 2\phi > 0$, $A_1 > 0$ and $c_0 > 0$.
(b) $B \in \mathcal{M}_p(\alpha, A_0)$ with $\alpha = 2 - 2\phi > 1$ and $A_0 > 0$.
(c) Conditions RCA and RCB hold with $\kappa = 2 - 2\phi > 0$ and $c_1 > 0$.

To show these claims, we need some lemmas and results in elementary calculus. In detail, first, we have

$$|f'(\omega)| \leq C|\omega|^{-(2\phi+1)}, \qquad |f''(\omega)| \leq C|\omega|^{-(2\phi+2)}. \tag{5.81}$$

For a proof of (5.81), we rewrite $f(\omega) = f^*(\omega)/|2\sin(\omega/2)|^{2\phi}$, where by assumption $f^*(\omega)$ is a continuous function that is twice differentiable except at 0, and $|(f^*)''(\omega)| \leq C|\omega|^{-2}$. It can be derived from basic properties in analysis that

$$|(f^*)''(\omega)| \leq C|\omega|^{-2}, \quad |(f^*)'(\omega)| \leq C|\omega|^{-1}, \quad \text{and} \quad |f^*(\omega)| \leq C. \tag{5.82}$$

At the same time, by elementary calculation,

$$\begin{aligned} |f'(\omega)| &\leq C|\omega|^{-(2\phi+1)}(|f^*(\omega)| + |\omega(f^*)'(\omega)|), \\ |f''(\omega)| &\leq C|\omega|^{-(2\phi+2)}(|f^*(\omega)| + |\omega(f^*)'(\omega)| + |\omega^2(f^*)''(\omega)|), \end{aligned}$$

and (5.81) follows by plugging in (5.82).

Second, we need the following lemma, whose proof is a simple exercise of analysis and omitted.

LEMMA 5.4. *Suppose $g$ is a symmetric real function which is differentiable in $[-\pi, 0) \cup (0, \pi]$ and $|g'(\omega)| \leq C|\omega|^{-\alpha}$ for some $\alpha \in (1, 2)$. Then as $x \to \infty$, $\int_{-\pi}^{\pi} \cos(\omega x) g(\omega) d\omega = O\left(|x|^{-(2-\alpha)}\right)$.*

We now show (a)-(c). Consider (a) first. First, by (5.81) and Lemma 5.4, $\int_{-\pi}^{\pi} \cos(k\omega) f(\omega) d\omega \leq Ck^{-(1-2\phi)}$ for large $k$, so that $|G(i,j)| \leq C(1 + |i - j|)^{-(1-2\phi)}$. Second, by well-known results on Toeplitz matrices, $\lambda_{\min}(G) \geq \min_{\omega \in [-\pi,\pi]} f(\omega) > 0$. Combining these, (a) holds with $\gamma = 1 - 2\phi$ and $c_0 = \min_{\omega \in [-\pi,\pi]} f(\omega)$.

Next, we consider (b). Recall that $B = DG$ where $D$ is the first-order row-differencing matrix. So $B(i,j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [\cos(k\omega) - \cos((k+1)\omega)] f(\omega) d\omega$, where $k = i - j$. Without loss of generality, we only consider the case $k \geq 1$.

Denote $g(\omega) = \omega f(\omega)$. By Fubini's theorem and integration by part,

$$\begin{aligned} B(i,j) &= \frac{1}{\pi}\int_0^{\pi}\left[\int_k^{k+1}\omega\sin(\omega x)dx\right]f(\omega)d\omega \\ &= \frac{1}{\pi}\int_k^{k+1}\left[\int_0^{\pi}g(\omega)\sin(\omega x)d\omega\right]dx \\ &= \frac{1}{\pi}\int_k^{k+1}\left[-g(\pi)\frac{\cos(\pi x)}{x}+\int_0^{\pi}\frac{\cos(\omega x)}{x}g'(\omega)d\omega\right]dx \\ &= -\frac{g(\pi)}{\pi}\int_k^{k+1}\frac{\cos(\pi x)}{x}dx+\frac{1}{2\pi}\int_k^{k+1}\frac{1}{x}\left[\int_{-\pi}^{\pi}\cos(\omega x)g'(\omega)d\omega\right]dx \\ &\equiv I_1+I_2 \end{aligned}$$

First, using integration by part, $|I_1| = \left|\pi^{-1}g(\pi)\int_k^{k+1}\frac{\sin(\pi x)}{\pi x^2}dx\right| = O(k^{-2})$. Second, similar to (5.81), we derive that $g''(\omega) = O(|\omega|^{-(1+2\phi)})$. Applying Lemma 5.4 to $g'$, we have $|\int_{-\pi}^{\pi}\cos(\omega x)g'(\omega)d\omega| \leq C|x|^{-(1-2\phi)}$, and so $|I_2| \leq \int_k^{k+1}Cx^{-(2-2\phi)}dx = O(k^{-(2-2\phi)})$. Combining these gives $|B(i,j)| \leq C(1+|i-j|)^{-(2-2\phi)}$, and (b) holds with $\alpha = 2-2\phi$.

Last, we show (c). Since $\varphi_\eta(z) = 1-z$, RCA holds trivially, and all remains is to check that RCB holds. Recall that $H = DGD'$, where $D$ is the first-order row-differencing matrix. The goal is to show there exsits a constant $c_1 > 0$ such that for any triplet $(k, b, V)$,

$$b'H^{V,V}b \geq c_1 k^{-(2-2\phi)}\|b\|^2, \tag{5.83}$$

where $1 \leq k \leq p$ is an integer, $b \in \mathbb{R}^k$ is a vector, and $V \subset \{1, 2, \ldots, p\}$ is a subset with $|V| = k$.

Towards this end, we introduce $f_1(\omega) = 4\sin^2(\omega/2)f(\omega)$, where we recall that $f$ is the spectral density associated with $G$. Fixing a triplet $(k, b, V)$, we write $b = (b_1, b_2, \ldots, b_k)'$ and $V = \{j_1, \cdots, j_k\}$ such that $j_1 < j_2 < \ldots < j_k$. By definitions and basic algebra,

$$\begin{aligned} H(i,j) &= G(i,j)-G(i+1,j)-G(i,j+1)+G(i+1,j+1) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[2\cos(k\omega)-\cos((k+1)\omega)-\cos((k-1)\omega)\right]f(\omega)d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(k\omega)f_1(\omega)d\omega, \qquad \text{where for short } k=i-j, \end{aligned}$$

which, together with direct calculations, implies that

$$b'H^{V,V}b = \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{s=1}^{k}\sum_{t=1}^{k}b_sb_t\cos((j_s-j_t)\omega)f_1(\omega)d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big|\sum_{s=1}^{k}b_se^{\sqrt{-1}j_s\omega}\Big|^2f_1(\omega)d\omega.$$

At the same time, note that $f_1(\omega) \geq C|\omega|^{2-2\phi}$ for any $\omega \neq 0$ and $|\omega| \leq \pi$. Combining these with symmetry and monotonicity gives

(5.84)
$$b'H^{V,V}b \geq \frac{C}{\pi}\int_0^\pi \big|\sum_{s=1}^k b_s e^{\sqrt{-1}j_s\omega}\big|^2 \omega^{2-2\phi} d\omega \geq \frac{C}{\pi}\int_{\pi/(2k)}^\pi \big|\sum_{s=1}^k b_s e^{\sqrt{-1}j_s\omega}\big|^2 \omega^{2-2\phi} d\omega.$$

Next, we write

$$\|b\|^2 = \frac{1}{\pi}\int_0^\pi \big|\sum_{s=1}^k b_s e^{\sqrt{-1}j_s\omega}\big|^2 d\omega = I + II, \tag{5.85}$$

where $I$ and $II$ are the integration in the interval of $[0, \pi/(2k)]$ and $[\pi/(2k), \pi]$, respectively. By (5.84) and the monotonicity of the function $\omega^{2-2\phi}$ in $[\pi/(2k), \pi]$,

$$b'H^{V,V}b \geq Ck^{-(2-2\phi)} \cdot \frac{1}{\pi}\int_{\pi/(2k)}^\pi \big|\sum_{s=1}^k b_s e^{\sqrt{-1}j_s\omega}\big|^2 d\omega \equiv Ck^{-(2-2\phi)} \cdot II. \tag{5.86}$$

At the same time, by the Cauchy-Schwartz inequality, $|\sum_{s=1}^k b_s e^{\sqrt{-1}j_s\omega}|^2 \leq (\sum_{s=1}^k |e^{\sqrt{-1}j_s\omega}|^2)(\sum_{s=1}^k |b_s|^2) = k\|b\|^2$, and so $I \leq \frac{1}{\pi}\int_0^{\pi/(2k)} k\|b\|^2 d\omega \leq \|b\|^2/2$. Inserting this into (5.85) gives

$$II \geq \|b\|^2 - \|b\|^2/2 = \|b\|^2/2, \tag{5.87}$$

and (5.83) follows by combining (5.86) and (5.87). □

5.5. *Proof of Lemma 1.5.* First, we show $r^*_{lts}(\vartheta) \equiv r^*_{lts}(\vartheta; f)$ is a decreasing function of $\vartheta$. Similarly to the proof of Theorem 1.3, in the definition of $\omega(F, N)$ and $\psi(F, N)$ (recall (1.33) and (1.34)), replace $G$ by $G^\infty$, and denote the new terms by $\omega^\infty(F, N) \equiv \omega^\infty(F, N; \vartheta, r, f)$ and $\psi^\infty(F, N) \equiv \psi^\infty(F, N; \vartheta, r, f)$, respectively. By similar argument in Lemma 5.3,

$$\rho^*_{lts}(\vartheta, r; f) = \min_{(F,N): F\cap N=\emptyset, F\neq\emptyset} \psi^\infty(F, N).$$

For each pair of sets $(F, N)$ and $\vartheta \in (0, 1)$ let $r^*(\vartheta; F, N) \equiv r^*(\vartheta; F, N, f)$ be the minimum $r$ such that $\psi^\infty(F, N; \vartheta, r, f) \geq 1$. It follows that

$$r^*_{lts}(\vartheta) = \max_{(F,N): F\cap N=\emptyset, F\neq\emptyset} r^*(\vartheta; F, N).$$

It is easy to see that $r^*(\vartheta; F, N)$ is a decreasing function of $\vartheta$ for each fixed $(F, N)$. So $r^*_{lts}(\vartheta)$ is also a decreasing function of $\vartheta$.

Next, we consider $\lim_{\vartheta\to 1} r^*_{lts}(\vartheta)$. In the special case of $F = \{j\}$ and $N = \emptyset$, $\omega^\infty(F,N) = 1$, $\lim_{\vartheta\to 1} r^*(\vartheta; F,N) = 1$, and so $\liminf_{\vartheta\to 1} r^*_{lts}(\vartheta) \geq 1$. At the same time, for any $(F,N)$ such that $|F|+|N| > 1$, $\psi^\infty(F,N) \geq \vartheta$ and so $\lim_{\vartheta\to 1} r^*(\vartheta; F,N) \leq 1$. Hence, $\limsup_{\vartheta\to 1} r^*_{lts}(\vartheta) \leq 1$. Combining these gives the claim.

Last, we consider $\lim_{\vartheta\to 0} r^*_{lts}(\vartheta)$. First, since $\lim_{\vartheta\to 0} \psi^\infty(F,N) = \omega^\infty(F,N)r/4$ for any fixed $(F,N)$, we have

$$\lim_{\vartheta\to 0} r^*_{lts}(\vartheta) = 4\Big[\min_{(F,N):F\cap N=\emptyset, F\neq\emptyset} \omega^\infty(F,N)\Big]^{-1}. \tag{5.88}$$

Second, by definitions,

$$\min_{(F,N):F\cap N=\emptyset, F\neq\emptyset} \omega^\infty(F,N) = \lim_{p\to\infty} \min_{(F,N):(F\cup N)\subset\{1,\cdots,p\}, F\cap N=\emptyset, F\neq\emptyset} \omega(F,N), \tag{5.89}$$

whenever the limit on the right hand side exists.

Third, note that (a) Given $F$, $\omega(F,N)$ decreases as $N$ increases and (b) Given $F\cup N$, $\omega(F,N)$ decreases as $N$ increases (the proofs are straightforward and we omit them). As a result, for all $(F,N)$ such that $(F\cup N) \subset \{1,\cdots,p\}$, $\omega(F,N)$ is minimized at $F = \{j\}$ and $N = \{1,\cdots,p\}\backslash\{j\}$ for some $j$, with the minimal value equaling the reciprocal of the $j$-th diagonal of $G^{-1}$. In other words,

$$\lim_{p\to\infty} \min_{(F,N):(F\cup N)\subset\{1,\cdots,p\}, F\cap N=\emptyset, F\neq\emptyset} \omega(F,N) = \big[\lim_{p\to\infty} \max_{1\leq j\leq p} G^{-1}(j,j)\big]^{-1}. \tag{5.90}$$

Fourth, if we write $G = G_p$ to emphasize on the size of $G$, then by basic algebra and the Toeplitz structure of $G$, we have $(G_p^{-1})(j,j) \leq (G_{p+k}^{-1})(j+k, j+k)$ for all $1 \leq k \leq p-j$ and $(G_p^{-1})(j,j) \leq (G_{p+k}^{-1})(j-k, j-k)$ for $1 \leq k \leq j-1$. Especially, if we take $k = \log(p)$, then it follows that

$$\lim_{p\to\infty} \max_{1\leq j\leq p} G^{-1}(j,j) = \lim_{p\to\infty} \max_{\log(p)\leq j\leq p-\log(p)} G^{-1}(j,j). \tag{5.91}$$

Last, we have the following lemma which is proved in Appendix A.

LEMMA 5.5. *Under conditions of Lemma 1.5,*

$$\lim_{p\to\infty} \max_{\log(p)\leq j\leq p-\log(p)} G^{-1}(j,j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f^{-1}(\omega)d\omega.$$

Combining (5.88)-(5.91) and using Lemma 5.5,

$$\lim_{\vartheta\to 0} r^*_{lts}(\vartheta) = 4\cdot\big[\lim_{p\to\infty} \max_{\log(p)\leq j\leq p-\log(p)} G^{-1}(j,j)\big] = \frac{2}{\pi}\int_{-\pi}^{\pi} f^{-1}(\omega)d\omega.$$

□

5.6. *Proof of Theorem 1.4.* Write for short $\hat{\beta} = \hat{\beta}^{case}$ and $\rho^*_{cp} = \rho^*_{cp}(\vartheta, r)$. It suffices to show

$$\text{Hamm}^*_p(\vartheta, r, G) \geq L_p p^{1-\rho^*_{cp}}; \tag{5.92}$$

and for any $\mu \in \Theta^*_p(\tau_p, a)$,

$$H_p(\hat{\beta}; \epsilon_p, \mu, G) \equiv \sum_{j=1}^{p} P\big(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j)\big) \leq L_p p^{1-\rho^*_{cp}} + o(1). \tag{5.93}$$

First, we show (5.92). The statement is similar to that of Theorem 1.1, but $d_p(\mathcal{G}^\diamond) \leq L_p$ does not hold. Therefore, we introduce a different graph $\mathcal{G}^\triangledown$ as follows: Define a counter part of $\rho^*_j(\vartheta, r, G)$ as

$$\tilde{\rho}^*_j(\vartheta, r, G) = \min_{(V_0, V_1):\min(V_0 \cup V_1) = j} \rho(V_0, V_1), \tag{5.94}$$

where $\min(V_0 \cup V_1) = j$ means $j$ is the smallest node in $V_0 \cup V_1$. Let $(V^*_{0j}, V^*_{1j})$ be the minimizer of (5.94), and when there is a tie, pick the one that appears first lexicographically. Define the graph $\mathcal{G}^\triangledown$ with nodes $\{1, \cdots, p\}$, and that there is an edge between nodes $j$ and $k$ whenever $(V^*_{0j} \cup V^*_{1j}) \cap (V^*_{0k} \cup V^*_{1k}) \neq \emptyset$.

Denote $d_p(\mathcal{G}^\triangledown)$ the maximum degree of nodes in $\mathcal{G}^\triangledown$. Similar to Theorem 1.1, as $p \to \infty$,

$$\text{Hamm}^*_p(\vartheta, r, G) \geq L_p [d_p(\mathcal{G}^\triangledown)]^{-1} \sum_{j=1}^{p} p^{-\tilde{\rho}^*_j(\vartheta, r, G)}. \tag{5.95}$$

The proof is a trivial extension of [20, Theorem 1.1] and we omit it. Moreover, the following lemma is proved below.

LEMMA 5.6. *As $p \to \infty$, $\max_{\log(p) \leq j \leq p - \log(p)} |\tilde{\rho}^*_j(\vartheta, r, G) - \rho^*_{cp}(\vartheta, r)| = o(1)$, and $d_p(\mathcal{G}^\triangledown) \leq L_p$.*

Combining (5.95) with Lemma 5.6 gives (5.92).

Second, we show (5.93). The change-point model is an 'extreme' case and Theorem 1.2 does not apply directly. However, once we justify the following claims (a)-(c), (5.93) follows by similar arguments in Theorem 1.2.

(a) SS property:

$$\sum_{j=1}^{p} P\big(\beta_j \neq 0, j \notin \mathcal{U}^*_p\big) \leq L_p p^{1-\rho^*_{cp}} + o(1).$$

(b) SAS property: If we view $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$, there is a fixed integer $l_0 > 0$ such that with probability at least $1 - o(1/p)$, each component of $\mathcal{U}_p^*$ has a size $\leq l_0$.

(c) A counter part of Lemma 2.6: For any $\log(p) \leq j \leq p - \log(p)$, and fixed $\mathcal{I} \trianglelefteq \mathcal{G}^+$ such that $j \in \mathcal{I}$ and $|\mathcal{I}| \leq l_0$, suppose we construct $\{\mathcal{I}^{(k')}, \mathcal{I}^{(k'),pe}, 1 \leq k' \leq N\}$ using the process introduced in the $PE$-step, and $j \in \mathcal{I}^{(k)}$. Then for any pair of sets $(V_0, V_1)$ such that $\mathcal{I}^{(k)} = V_0 \cup V_1$,

$$\rho_j(V_0, V_1; \mathcal{I}^{(k)}) \geq \rho_{cp}^* + o(1),$$

where $\rho_j(V_0, V_1; \mathcal{I}^{(k)})$ is defined in (2.59).

Consider (a) first. Following the proof of Lemma 2.1 until (5.119), we find that for each $\log(p) \leq j \leq p - \log(p)$,

$$P\big(\beta_j \neq 0, j \notin \mathcal{U}_p^*\big) \leq \sum_{(\mathcal{I},F,N): j \in \mathcal{I} \trianglelefteq \mathcal{G}^*, |\mathcal{I}| \leq m, F \cup N = \mathcal{I}, F \cap N = \emptyset, F \neq \emptyset} L_p p^{-|\mathcal{I}|\vartheta - [(\sqrt{\omega_0 r} - \sqrt{q})_+]^2} + L_p p^{-(m+1)\vartheta} + o(1/p)$$

where $\omega_0 = \tau_p^{-2}(\beta^F)'[Q^{F,F} - Q^{F,N}(Q^{N,N})^{-1}Q^{N,F}]\beta^F$ and $Q$ is defined as in (1.15). First, by the choice of $m$, $L_p p^{-(m+1)\vartheta} \leq L_p p^{-\rho_{cp}^*}$. Second, using similar arguments in Lemma 2.1, the summation contains at most $L_p$ terms. Third, by (1.35), $\omega_0 \geq \tilde{\omega}(F, N)$. Combining the above, it suffices to show for each triplet $(\mathcal{I}, F, N)$ in the summation,

$$|\mathcal{I}|\vartheta + [(\sqrt{\tilde{\omega}(F,N)r} - \sqrt{q})_+]^2 \geq \rho_{cp}^*. \tag{5.96}$$

The key to (5.96) is to show

$$\tilde{\omega}(F, N) \geq 1/2. \tag{5.97}$$

Once (5.97) is proved, since $q \leq \frac{r}{4}(\sqrt{2} - 1)^2$,

$$|\mathcal{I}|\vartheta + [(\sqrt{\tilde{\omega}(F,N)r} - \sqrt{q})_+]^2 \geq |\mathcal{I}|\vartheta + r/4 \geq \rho_{cp}^*,$$

where in the last inequality we use the facts $\rho_{cp}^* \leq \vartheta + r/4$ and $|\mathcal{I}| \geq 1$. This gives (5.96).

All remains is to show (5.97). We argue that it suffices to consider those $(\mathcal{I}, F, N)$ where both $\mathcal{I}(= F \cup N)$ and $F$ are formed by consecutive nodes. First, since $G$ is tri-diagonal, the definition of $\mathcal{G}^*$ implies that any $\mathcal{I} \trianglelefteq \mathcal{G}^*$ is formed by consecutive nodes. Second, by (1.35) and basic algebra,

$$\tilde{\omega}(F, N) = \min_{\xi \in \mathbb{R}^{|F|}: |\xi_i| \geq 1} \xi'[(Q^{-1})^{F,F}]^{-1}\xi, \tag{5.98}$$

where $Q$ is defined in (1.15). Note that $B$ is an identity matrix and $\mathcal{I}^{ps} = \mathcal{I}$. So $Q^{-1} = H^{\mathcal{I},\mathcal{I}}$, which is a tri-diagonal matrix. It follows from (5.98) that if $F$ is not formed by consecutive nodes, there exist $F_1 \subset F$ and $N_1 = \mathcal{I} \backslash F_1$ such that $\tilde{\omega}(F_1, N_1) \leq \tilde{\omega}(F, N)$. The argument then follows.

From now on, we focus on $(\mathcal{I}, F, N)$ such that both $\mathcal{I}$ and $F$ are formed by consecutive nodes. Elementary calculation yields

$$[(Q^{-1})^{F,F}]^{-1} = (H^{F,F})^{-1} = \Omega^{(k)} - \frac{1}{k+1}\eta\eta', \tag{5.99}$$

where $k = |F|$, $\Omega^{(k)}$ is the $k \times k$ matrix defined by $\Omega^{(k)}(i,j) = \min\{i,j\}$ and $\eta = (1, \cdots, k)'$. We see that $\tilde{\omega}(F, N)$ only depends on $k$. When $k = 1$, $\tilde{\omega}(F, N) = 1/2$ by direct calculations following (5.98) and (5.99). When $k \geq 2$, from (5.98) and (5.99),

$$\tilde{\omega}(F, N) = \min_{\xi \in \mathbb{R}^{|F|}: |\xi_i| \geq 1} \left[ \sum_{l=1}^{k} (\xi_l + \cdots + \xi_k)^2 - \frac{1}{k+1}(\xi_1 + 2\xi_2 + \cdots + k\xi_k)^2 \right].$$

Let $s_l = \sum_{j=l}^{k} \xi_j$. The above right hand side is lower bounded by $\sum_{l=1}^{k} s_l^2 - (\sum_{l=1}^{k} s_l)^2/k = \sum_{l<l'}(s_l - s_{l'})^2/k$, where $\sum_{l<l'}(s_l - s_{l'})^2 \geq \sum_{l=1}^{k-1}(s_{l+1} - s_l)^2 \geq k-1$. Therefore,

$$\tilde{\omega}(F, N) \geq (k-1)/k \geq 1/2.$$

This proves (5.97).

Next, consider (b). We check RCB, and the remaining proof is exactly the same as in Lemma 2.2. Towards this end, the goal is to show there exists a constant $c_1 > 0$ such that for any $(k, V)$ where $V \subset \{1, \cdots, p\}$ and $k = |V|$,

$$\lambda_{\min}(H^{V,V}) \geq c_1 k^{-2}. \tag{5.100}$$

Since $H$ is tri-diagonal, it suffices to show that (5.100) holds when $V$ is formed by consecutive nodes, i.e., $V = \{j, \cdots, j+k\}$ for some $1 \leq j \leq p-k$. In this case, we introduce a matrix $\Sigma^{(k)}$, which is 'smaller' than $H^{V,V}$ but much easier to analyse:

$$\Sigma^{(k)}(i,j) = 2 \cdot 1\{i = j\} - 1\{|i-j| = 1\} - 1\{i = j = k\}, \quad 1 \leq i, j \leq k.$$

It is easy to see that $H^{V,V} - \Sigma^{(k)}$ is positive semi-definite. Hence,

$$\lambda_{\min}(H^{V,V}) \geq \lambda_{\min}(\Sigma^{(k)}). \tag{5.101}$$

Observing that $(\Sigma^{(k)})^{-1} = \Omega^{(k)}$, where $\Omega^{(k)}$ is as in (5.99), we have

$$\lambda_{\min}(\Sigma^{(k)}) = \left[\lambda_{\max}(\Omega^{(k)})\right]^{-1} \geq \left[\|\Omega^{(k)}\|_\infty\right]^{-1} = 2/(k^2+k). \tag{5.102}$$

Combining (5.101)-(5.102) gives (5.100).

Finally, consider (c). Fix $1 \leq j \leq p$ and the triplet $(\mathcal{I}^{(k)}, V_0, V_1)$, where $|\mathcal{I}^{(k)}| \leq l_0$. The goal is to show

$$\rho_j(V_0, V_1; \mathcal{I}^{(k)}) \geq \rho_{cp}^* + o(1). \tag{5.103}$$

Introduce the following quantities: From the $PE$-step and the choice $\ell^{pe} = 2\log(p)$, we can write

$$\mathcal{I}^{(k),pe} = \{j_1+1, \cdots, j_1+L\} \quad \text{and} \quad \mathcal{I}^{(k)} = \{j_1+M_1, \cdots, j_1+M_2\},$$

where the integers $L$, $M_1$ and $M_2$ staisfy

$$M_2 - M_1 \leq l_0 + 1, \quad M_1 \geq [\log(p)]^{1/(l_0+1)}, \quad (L-M_2)/M_1 \geq [\log(p)]^{1/(l_0+1)}. \tag{5.104}$$

Denote $K = M_2 - M_1 + 1$, $M_0 = M_1 - \frac{M_1^2}{L+1}$ and $\mathcal{I}'' = \{M_0, \cdots, M_0+K-1\}$. Let $\mathcal{F}$ be the one-to-one mapping from $\mathcal{I}^{(k)}$ to $\mathcal{I}''$ such that $\mathcal{F}(i) = i - (j_1 + M_1) + M_0$. Denote $V_0'' = \mathcal{F}(V_0)$ and $V_1'' = \mathcal{F}(V_1)$. Recall the definitions of $\varpi_j(V_0, V_1; \mathcal{I}^{(k)})$ and $\varpi^*(V_0, V_1)$ (see (2.59) and (1.25)). We claim that

$$\varpi_j(V_0, V_1; \mathcal{I}^{(k)}) \geq \varpi^*(V_0'', V_1'') + o(1). \tag{5.105}$$

Once we have (5.105), plug it into the definition $\rho_j(V_0, V_1; \mathcal{I}^{(k)})$ and use the monotonicity of the function $f(x) = [(x - a/x)_+]^2$ over $(0, \infty)$ when $a > 0$. It follows that

$$\rho_j(V_0, V_1; \mathcal{I}^{(k)}) \geq \max\{|V_0|, |V_1|\}\vartheta + \frac{1}{4}\left[\left(\sqrt{\varpi^* r} - \frac{\big||V_1| - |V_0|\big|\vartheta}{\sqrt{\varpi^* r}}\right)_+\right]^2 + o(1).$$

where $\varpi^*$ is short for $\varpi^*(V_0'', V_1'')$. Compare the first term on the right hand side with (1.26) and recall that $|V_0''| = |V_0|$ and $|V_1''| = |V_1|$. It follows that

$$\rho_j(V_0, V_1; \mathcal{I}^{(k)}) \geq \rho(V_0'', V_1'') + o(1). \tag{5.106}$$

Moreover, since $M_0 = \min(V_0'', V_1'')$, by (5.94),

$$\rho(V_0'', V_1'') \geq \tilde{\rho}_{M_0}^*. \tag{5.107}$$

Note that (5.104) implies $M_0 \gtrsim M_1 \geq [\log(p)]^{1/(1+l_0)}$. By a trivial extension of Lemma 5.6, we can derive that $\max_{(\log(p))^{1/(1+l_0)} \leq j \leq p-(\log(p))^{1/(1+l_0)}} |\tilde{\rho}_j^* - \rho_{cp}^*| = o(1)$. These together imply

$$\tilde{\rho}_{M_0}^* = \rho_{cp}^* + o(1). \tag{5.108}$$

Combining (5.106)-(5.108) gives (5.103).

What remains is to show (5.105). The proof is similar to that of (5.152). In detail, write for short $\varpi_j = \varpi_j(V_0, V_1; \mathcal{I}^{(k)})$, $\varpi^* = \varpi^*(V_0'', V_1'')$, $B_1 = B^{\mathcal{I}^{(k)},pe,\mathcal{I}^{(k)}}$, $H_1 = H^{\mathcal{I}^{(k)},pe,\mathcal{I}^{(k)},pe}$ and $Q_1 = B_1' H_1^{-1} B_1$. By similar arguments in (5.153), $\varpi_j \geq \min_{j \in \mathcal{I}} \varpi_j$, and there exists a constant $a_1 > 0$ such that

$$|\min_{j\in\mathcal{I}} \varpi_j - \varpi^*| \leq \max_{\xi \in \mathbb{R}^K: \|\xi\|_\infty \leq 2a_1} \left|\xi'(G^{\mathcal{I}'',\mathcal{I}''} - Q_1)\xi\right| \leq C\|G^{\mathcal{I}'',\mathcal{I}''} - Q_1\|.$$

Therefore, it suffices to show that

$$\|G^{\mathcal{I}'',\mathcal{I}''} - Q_1\| = o(1). \tag{5.109}$$

Note that $Q_1$ is the $(\mathcal{I}', \mathcal{I}')$-block of $H_1^{-1}$, where the index set $\mathcal{I}' = \{M_1, \cdots, M_2\}$. By (5.99), $H_1^{-1} = \Omega^{(L)} - \frac{1}{L+1}\eta\eta'$, where $\eta = (1, 2, \cdots, L)'$. It follows that

$$Q_1 = (M_1 - 1)1_K 1_K' + \Omega^{(K)} - \frac{1}{L+1}\xi\xi',$$

where $1_K$ is the $K$-dimensional vector whose elements are all equal to 1, and $\xi = (M_1, \cdots, M_2)'$. Define the $L \times L$ matrix $\Delta$ by $\Delta(i,j) = \frac{ij - M_1^2}{L+1}$, for $1 \leq i, j \leq L$ and let $\Delta_1$ be the submatrix of $\Delta$ by restricting the rows and columns to $\mathcal{I}'$. By these notations,

$$Q_1 = (M_0 - 1)1_K 1_K' + \Omega^{(K)} - \Delta_1.$$

At the same time, we observe that

$$G^{\mathcal{I}'',\mathcal{I}''} = (M_0 - 1)1_K 1_K' + \Omega^{(K)}.$$

Combining the above yields that $G^{\mathcal{I}'',\mathcal{I}''} - Q_1 = \Delta_1$. Note that $|\Delta(i,j)| \leq \frac{M_2^2 - M_1^2}{L+1} \leq \frac{(l_0+1)(2M_1+l_0+1)}{L+1} = o(1)$ for all $i, j \in \mathcal{I}'$. Hence, $\|\Delta_1\| = o(1)$ and (5.109) follows directly. ∎

5.6.1. *Proof of Lemma 5.6.* To show the claim, we need to introduce some quantities and lemmas. First, by a trivial extension of Lemma 5.3,

$$\tilde{\rho}_j^*(\vartheta, r, G) = \min_{(F,N): \min(F \cup N) = j, F \cap N = \emptyset, F \neq \emptyset} \psi(F, N).$$

where $\psi(F,N) = \psi(F,N;\vartheta,r,G)$, defined in (1.33).

Second, let $\mathcal{R}_p$ denote the collection of all subsets of $\{1,\cdots,p\}$ that are formed by consecutive nodes. Define

$$\tilde{\tilde{\rho}}_j^*(\vartheta,r,G) = \min_{(F,N):\min(F\cup N)=j, F\cap N=\emptyset, F\neq\emptyset, F\cup N\in\mathcal{R}_p, F\in\mathcal{R}_p, |F|\leq 3, |N|\leq 2} \psi(F,N),$$

where we emphasize that the minimum is taken over finite pairs $(F,N)$. The following lemma is proved in Appendix A.

Lemma 5.7. *As $p\to\infty$, $\max_{\log(p)\leq j\leq p-\log(p)} |\tilde{\rho}_j^*(\vartheta,r,G) - \tilde{\tilde{\rho}}_j^*(\vartheta,r,G)| = o(1)$.*

Third, for each dimension $k$, define the $k\times k$ matrix $\Sigma_*^{(k)}$ as

$$\Sigma_*^{(k)}(i,j) = 2\cdot 1\{i=j\} - 1\{|i-j|=1\}, \tag{5.110}$$

except that $\Sigma_*^{(k)}(1,1) = \Sigma_*^{(k)}(k,k) = 1$, and the $k\times k$ matrix $\Omega_*^{(k)}$ as

$$\Omega_*^{(k)}(i,j) = \min\{i,j\} - 1. \tag{5.111}$$

Let

$$\omega^{(\infty)}(F,N) = \begin{cases} \min_{\xi\in\mathbb{R}^{|F|}:|\xi_j|\geq 1} \xi'[(\Sigma_*^{(k)})^{F,F}]^{-1}\xi, & |N|>0 \\ \min_{\xi\in\mathbb{R}^{|\mathcal{I}|}:|\xi_j|\geq 1, 1'\xi=0} \xi'\Omega_*^{(k)}\xi, & |N|=0,\ |F|>1 \\ \infty, & |N|=0,\ |F|=1 \end{cases}$$

and define $\psi^{(\infty)}(F,N) = \psi^{(\infty)}(F,N;\vartheta,r,G)$, a counter part of $\psi(F,N)$, by replacing $\omega(F,N)$ by $\omega^{(\infty)}(F,N)$ in the definition (1.33). Let

$$\rho^{(\infty)}(\vartheta,r) = \min_{(F,N):\min(F\cup N)=1, F\cap N=\emptyset, F\neq\emptyset, F\in\mathcal{R}_p, F\cup N\in\mathcal{R}_p, |F|\leq 3, |N|\leq 2} \psi^{(\infty)}(F,N),$$

where we note that $\rho^{(\infty)}(\vartheta,r)$ does not depend on $j$. The following lemma is proved in Appendix A.

Lemma 5.8. *As $p\to\infty$, $\max_{\log(p)\leq j\leq p-\log(p)} |\tilde{\tilde{\rho}}_j^*(\vartheta,r,G) - \rho^{(\infty)}(\vartheta,r)| = o(1)$.*

Now, we show the claims. Write for short $\tilde{\rho}_j^* = \tilde{\rho}_j^*(\vartheta,r,G)$, and $\tilde{\tilde{\rho}}_j^*$, $\rho_{cp}^*$ similarly. First, we show

$$d_p(\mathcal{G}^{\triangledown}) \leq L_p.$$

Denote $(F_j^*, N_j^*)$ the minimum in defining $\tilde{\tilde{\rho}}_j^*$, and if there is a tie, we pick the one that appears first lexicographically. By definition and Lemma 5.7, for any $\log(p) \leq j \leq p - \log(p)$,

$$\psi(F_j^*, N_j^*) \equiv \tilde{\tilde{\rho}}_j^* = \tilde{\rho}_j^* + o(1), \qquad \text{and} \qquad (F_j^* \cup N_j^*) \subset \{j, \cdots, j+4\}.$$

By the definition of $\mathcal{G}^{\nabla}$, these imply that there is an edge between nodes $j$ and $k$ only when $|k - j| \leq 4$. So $d_p(\mathcal{G}^{\nabla}) \leq C$.

Next, we show for all $\log(p) \leq j \leq p - \log(p)$,

$$\tilde{\rho}_j^* = \rho_{cp}^* + o(1).$$

By Lemma 5.7 and Lemma 5.8, it suffices to show

$$\rho^{(\infty)} = \rho_{cp}^*. \tag{5.112}$$

Introduce the function $\nu(\cdot; F, N)$ for each $(F, N)$:

$$\nu(x; F, N) = \begin{cases} (|F| + 2|N|)/2 + \omega^{(\infty)} x/4, & |F| \text{ is even}, \\ (|F| + 2|N| + 1)/2 + [(\sqrt{\omega^{(\infty)} x} - 1/\sqrt{\omega^{(\infty)} x})_+]^2/4, & |F| \text{ is odd}, \end{cases}$$

where $\omega^{(\infty)}$ is short for $\omega^{(\infty)}(F, N)$. Then we can write

$$\psi^{(\infty)}(F, N; \vartheta, r, G) = \vartheta \cdot \nu(r/\vartheta; F, N).$$

Let $\nu^*(x) = \min_{(F,N)} \nu(x; F, N)$, where the minimum is taken over those $(F, N)$ in defining $\rho^{(\infty)}$. It follows that

$$\rho^{(\infty)}(\vartheta, r) = \min_{(F,N)} \vartheta \cdot \nu(r/\vartheta; F, N) = \vartheta \cdot \nu^*(r/\vartheta). \tag{5.113}$$

Below, we compute the function $\nu^*(\cdot)$ by computing the functions $\nu(\cdot; F, N)$ for the finite pairs $(F, N)$ in defining $\rho^{(\infty)}$. After excluding some obviously non-optimal pairs, all possible cases are displayed in Table 5. Using Table 5, we can further exclude the cases with $|F| = 3$. In the remaining, for each fixed value of $\omega^{(\infty)}$, we keep two pairs of $(F, N)$ which minimize $|F| + 2|N|$ among those with $|F|$ odd and even respectively. The results are displayed in Table 6. Then $\nu^*(\cdot)$ is the lower envelope of the four functions listed. Direct calculations yield

$$\nu^*(x) = \begin{cases} 1 + x/4, & 0 < x \leq 6 + 2\sqrt{10}; \\ 3 + (\sqrt{x} - 2/\sqrt{x})^2/8, & x > 6 + 2\sqrt{10}. \end{cases}$$

Plugging this into (5.113) and comparing it with the definition of $\rho_{cp}^*$, we obtain (5.112). □

TABLE 5
*Calculation of* $\omega^{(\infty)}(F,N)$

| $F$ | $N$ | $(\Sigma_*^{(k)})^{F,F}$ | $\Omega_*^{(k)}$ | $\xi^*$ | $\omega^{(\infty)}(F,N)$ |
|---|---|---|---|---|---|
| $\{1\}$ | $\{2\}$ | 1 | - | 1 | 1 |
| $\{2\}$ | $\{1,3\}$ | 2 | - | 1 | $\frac{1}{2}$ |
| $\{1,2\}$ | $\emptyset$ | - | $\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}$ | $(1,-1)'$ | 1 |
| $\{1,2\}$ | $\{3\}$ | $\begin{pmatrix}1 & -1\\ -1 & 2\end{pmatrix}$ | - | $(1,-1)'$ | 1 |
| $\{2,3\}$ | $\{1,4\}$ | $\begin{pmatrix}2 & -1\\ -1 & 2\end{pmatrix}$ | - | $(1,-1)'$ | $\frac{2}{3}$ |
| $\{1,2,3\}$ | $\emptyset$ | - | $\begin{pmatrix}0 & 0 & 0\\ 0 & 1 & 1\\ 0 & 1 & 2\end{pmatrix}$ | $(1,-2,1)'$ | 2 |
| $\{1,2,3\}$ | $\{4\}$ | $\begin{pmatrix}1 & -1 & 0\\ -1 & 2 & -1\\ 0 & -1 & 2\end{pmatrix}$ | - | $(1,-\frac{3}{2},1)'$ | $\frac{3}{2}$ |
| $\{2,3,4\}$ | $\{1,5\}$ | $\begin{pmatrix}2 & -1 & 0\\ -1 & 2 & -1\\ 0 & -1 & 2\end{pmatrix}$ | - | $(1,-1,1)'$ | 1 |

TABLE 6
*Calculation of* $\nu(x;F,N)$

| $\omega^{(\infty)}$ | $\lvert F\rvert$ | $\lvert N\rvert$ | $\nu(x;F,N)$ | $\\|\xi^*\\|_\infty$ |
|---|---|---|---|---|
| 1 | 1 | 1 | $2+\frac{1}{4}(\sqrt{x}-\frac{1}{\sqrt{x}})_+^2$ | 1 |
| 1 | 2 | 0 | $1+\frac{x}{4}$ | 1 |
| $\frac{1}{2}$ | 1 | 2 | $3+\frac{1}{8}(\sqrt{x}-\frac{2}{\sqrt{x}})_+^2$ | 1 |
| $\frac{2}{3}$ | 2 | 2 | $3+\frac{x}{6}$ | 1 |

5.7. *Proof of Lemma 2.1.* Fix $\vartheta$ and $r$. Write for short $\rho_j^* = \rho_j^*(\vartheta,r,G)$. To show the claim, it suffices to show for each $1 \le j \le p$,

$$P\big(\beta_j \neq 0,\ j \notin \mathcal{U}_p^*\big) \le L_p[p^{-\rho_j^*} + p^{-(m+1)\vartheta}] + o(1/p). \tag{5.114}$$

Fix $1 \le j \le p$. Recall that $\mathcal{G}_S^*$ is the subgraph of $\mathcal{G}^*$ by restricting the nodes into $S(\beta)$. Over the event $\{\beta_j \neq 0\}$, there is a unique component $\mathcal{I}$ such that $j \in \mathcal{I} \lhd \mathcal{G}_S^*$. By [14, 20], $|\mathcal{I}| \le m$ except for a probability of at most $L_p p^{-(m+1)\vartheta}$, where the randomness comes from the law of $\beta$. Denote this event as $A_p = A_{p,j}$. To show (5.114), it suffices to show

$$P\big(\beta_j \neq 0,\ j \notin \mathcal{U}_p^*,\ A_p\big) \le L_p p^{-\rho_j^*} + o(1/p). \tag{5.115}$$

Note that $\mathcal{I}$ depends on $\beta$ (and so is random), and also that over the event $A_p$, any realization of $\mathcal{I}$ is a connected subgraph in $\mathcal{G}^*$ with size $\le m$.

Therefore,

$$P\big(\beta_j \neq 0,\ j \notin \mathcal{U}_p^*,\ A_p\big) \leq \sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^*, |\mathcal{I}| \leq m} P\big(j \in \mathcal{I} \lhd \mathcal{G}_S^*,\ j \notin \mathcal{U}_p^*,\ A_p\big),$$

where on the right hand side, we have misused the notation slightly by denoting $\mathcal{I}$ as a fixed (non-random) connected subgraph of $\mathcal{G}^*$. Since $\mathcal{G}^*$ is $K_p$-sparse (see Lemma 5.1), for any fixed $j$, there are no more than $C(eK_p)^m$ connected subgraph $\mathcal{I}$ such that $j \in \mathcal{I}$ and $|\mathcal{I}| \leq m$ [14]. Noticing that $C(eK_p)^m \leq L_p$, to show (5.115), it is sufficient to show for any fixed $\mathcal{I}$ such that $j \in \mathcal{I} \trianglelefteq \mathcal{G}^*$ and $|\mathcal{I}| \leq m$,

$$P\big(j \in \mathcal{I} \lhd \mathcal{G}_S^*,\ j \notin \mathcal{U}_p^*,\ A_p\big) \leq L_p p^{-\rho_j^*} + o(1/p). \tag{5.116}$$

Fix such an $\mathcal{I}$. The subgraph (as a whole) has been screened in some sub-stage of the $PS$-step, say, sub-stage $t$. Let $\hat{N} = \mathcal{U}^{(t-1)} \cap \mathcal{I}$ and $\hat{F} = \mathcal{I} \backslash \hat{N}$ be as in the initial sub-step of the PS-step. By definitions, the event $\{j \notin \mathcal{U}_p^*\}$ is contained in the event that $\mathcal{I}$ fails to pass the $\chi^2$-test in (1.17). As a result,

$$\begin{aligned} P\big(j \in \mathcal{I} \lhd \mathcal{G}_S^*, j \notin \mathcal{U}_p^*, A_p\big) &\leq P\big(j \in \mathcal{I} \lhd \mathcal{G}_S^*, T(d, \hat{F}, \hat{N}) \leq 2q(\hat{F}, \hat{N}) \log(p), A_p\big) \\ &\leq \sum_{(F,N): F \cup N = \mathcal{I}, F \cap N = \emptyset, F \neq \emptyset} P\big(j \in \mathcal{I} \lhd \mathcal{G}_S^*, T(d, F, N) \leq 2q(F, N) \log(p), A_p\big), \end{aligned}$$

where $(F, N)$ are fixed (non-random) subsets, and $q = q(F, N)$ is either as in (1.31) or in (1.36). Since $|\mathcal{I}| \leq m$, the summation in the second line only involves at most finite terms. Therefore, to show (5.116), it suffices to show for each fixed triplet $(\mathcal{I}, F, N)$ satisfying $j \in \mathcal{I} \trianglelefteq \mathcal{G}^*$, $|\mathcal{I}| \leq m$, $F \cup N = \mathcal{I}$, $F \cap N = \emptyset$ and $F \neq \emptyset$,

$$P\big(j \in \mathcal{I} \lhd \mathcal{G}_S^*,\ T(d, F, N) \leq 2q(F, N) \log(p),\ A_p\big) \leq L_p p^{-\rho_j^*} + o(1/p). \tag{5.117}$$

Now, we show (5.117). The following lemma is proved below.

Lemma 5.9. *For each fixed $(\mathcal{I}, F, N)$ such that $\mathcal{I} = F \cup N$, $F \cap N = \emptyset$, $F \neq \emptyset$ and $|\mathcal{I}| \leq m$, there exists a random variable $T_0$ such that with probability at least $1 - o(1/p)$, $|T(d, F, N) - T_0| \leq C(\log(p))^{1/\alpha}$, and conditioning on $\beta^{\mathcal{I}}$, $T_0$ has a non-central $\chi^2$-distribution with the degree of freedom $k \leq |\mathcal{I}|$ and the non-centrality parameter*

$$\delta_0 = (\beta^F)' \left[Q^{F,F} - Q^{F,N}(Q^{N,N})^{-1} Q^{N,F}\right] \beta^F,$$

*where $Q$ is as defined in (1.15).*

Fix a triplet $(\mathcal{I}, F, N)$ and let $\delta_0$ be as in Lemma 5.9. Then

$$\begin{aligned}(5.118)\ & P\big(\mathcal{I} \lhd \mathcal{G}_S^*,\ T(d,F,N) \le 2q(F,N)\log(p),\ A_{p,j}\big) \\ &\le P\big(\mathcal{I} \lhd \mathcal{G}_S^*,\ T_0 \le 2q(F,N)\log(p) + C(\log(p))^{1/\alpha}\big) + o(1/p) \\ &\le P\big(\mathcal{I} \lhd \mathcal{G}_S^*\big) \cdot P\big(T_0 \le 2q(F,N)\log(p) + C(\log(p))^{1/\alpha}\ \big|\beta^{\mathcal{I}}\big) + o(1/p).\end{aligned}$$

Denote $\omega_0 = \tau_p^{-2}\delta_0$. By Lemma 5.9, $(T_0|\beta^{\mathcal{I}}) \sim \chi_k^2(2r\omega_0\log(p))$, where $k \le m$. In addition, $(\log(p))^{1/\alpha} \ll \log(p)$ by recalling that $\alpha > 1$. Combining these and using the basic property of non-central $\chi^2$-distributions,

$$P\big(T_0 \le 2q(F,N)\log(p) + C(\log(p))^{1/\alpha}\ \big|\beta^{\mathcal{I}}\big) \le L_p p^{-[(\sqrt{\omega_0 r} - \sqrt{q(F,N)})_+]^2}.$$

Inserting this into (5.118) and noting that $P\big(\mathcal{I} \lhd \mathcal{G}_S^*\big) \le L_p p^{-|\mathcal{I}|\vartheta}$, we have

$$P\big(\mathcal{I}\lhd\mathcal{G}_S^*,\ T(d,F,N) \le 2q(F,N)\log(p),\ A_p\big) \le L_p p^{-|\mathcal{I}|\vartheta - [(\sqrt{\omega_0 r} - \sqrt{q(F,N)})_+]^2} + o(1/p).$$

Comparing this with (5.117) and using the expression of $\rho_j^*$ in Lemma 5.3, it suffices to show

$$|\mathcal{I}|\vartheta + [(\sqrt{\omega_0 r} - \sqrt{q(F,N)})_+]^2 \ge \psi(F,N). \tag{5.119}$$

Recall that $q = q(F,N)$ is chosen from either (1.31) or (1.36). In the former case, since $\omega_0 \ge \tilde{\omega}(F,N)$ by definition (see (1.35)), it follows immediately from (1.31) and (1.32) that (5.119) holds. Therefore, we only consider the latter, in which case $q(F,N) = \tilde{q}|F|$ and (5.119) reduces to

$$|\mathcal{I}|\vartheta + [(\sqrt{\omega_0 r} - \sqrt{\tilde{q}|F|})_+]^2 \ge \psi(F,N). \tag{5.120}$$

By the expression of $\psi(F,N)$,

$$\psi(F,N) \le (|\mathcal{I}| - |F|/2)\vartheta + (\omega r/4 + \vartheta/2) \le |\mathcal{I}|\vartheta + \omega r/4,$$

where $\omega$ is a shorthand of $\omega(F,N)$. Therefore, to show (5.120), it suffices to check

$$\big(\sqrt{\omega_0 r} - \sqrt{\tilde{q}|F|}\ \big)_+ \ge \sqrt{\omega r}/2. \tag{5.121}$$

Towards this end, recalling that $F \subset \mathcal{I}$, we let $\Sigma$ and $\tilde{\Sigma}$ be the respective submatrices of $(G^{\mathcal{I},\mathcal{I}})^{-1}$ and $Q^{-1}$ formed by restricting the rows and columns from $\mathcal{I}$ to $F$. Let $\xi^* = \tau_p^{-1}\beta^F$. By elementary calculation and noting that $a > a_g^*(G)$,

$$\omega = \min_{\xi \in \mathbb{R}^{|F|}: 1 \le |\xi_i| \le a} \xi'\Sigma^{-1}\xi, \qquad \omega_0 = (\xi^*)'\tilde{\Sigma}^{-1}\xi^*.$$

On one hand, since $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$ and $|\mathcal{I}| \leq m \leq g$,

$$\omega \geq |F| \cdot \lambda_{\min}(G^{\mathcal{I},\mathcal{I}}) \geq c_0 \cdot |F|. \tag{5.122}$$

On the other hand, noting that $\|\xi^*\|_\infty \leq a$,

$$|\omega - \omega_0| \leq \max_{\xi \in \mathbb{R}^{|F|}: 1 \leq |\xi_i| \leq a} |\xi'(\Sigma^{-1} - \tilde{\Sigma}^{-1})\xi| \leq (a^2 \cdot \|\Sigma^{-1} - \tilde{\Sigma}^{-1}\|)|F|. \tag{5.123}$$

We argue that $\|\Sigma^{-1} - \tilde{\Sigma}^{-1}\|$ can be taken to be sufficiently small by $\ell^{ps}$ sufficiently large. To see the point, note that $\|\Sigma^{-1} - \tilde{\Sigma}^{-1}\| \leq \|G^{\mathcal{I},\mathcal{I}}\|^2 \|Q^{-1}\|^2 \|G^{\mathcal{I},\mathcal{I}} - Q\|$. First, since $|\mathcal{I}| \leq m$, $\|G^{\mathcal{I},\mathcal{I}}\|^2 \leq C$. Second, note that $Q$ is the Fisher Information Matrix associated with the model $d^{\mathcal{I}^{ps}} \sim N(B^{\mathcal{I}^{ps},\mathcal{I}}\beta^{\mathcal{I}}, H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})$. Using Lemma 1.2 and (5.143), $\|G^{\mathcal{I},\mathcal{I}} - Q\| \leq C(\ell^{ps})^{-\gamma}$. Third, $\|(G^{\mathcal{I},\mathcal{I}})^{-1}\| \leq c_0^{-1}$, since $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$. Finally, $\|Q^{-1}\| \leq 2c_0^{-1}$ when $G^{\mathcal{I},\mathcal{I}}$ and $Q$ are sufficiently close. Combining these gives that $\|\Sigma^{-1} - \tilde{\Sigma}^{-1}\| \leq C(\ell^{ps})^{-\gamma}$, for sufficiently large $\ell^{ps}$, and the claim follows.

As a result, by taking $\ell^{ps}$ a sufficiently large constant integer, we have

$$a^2 \|\Sigma^{-1} - \tilde{\Sigma}^{-1}\| \leq \big(\frac{1}{2}\sqrt{c_0} - \sqrt{\tilde{q}/r}\big)^2, \tag{5.124}$$

where we note the right hand side is a fixed positive constant. Combining (5.122)-(5.124),

$$\sqrt{(\omega - \omega_0)_+} \leq \big(\frac{1}{2}\sqrt{c_0} - \sqrt{\tilde{q}/r}\big)\sqrt{|F|} \leq \frac{1}{2}\sqrt{\omega} - \sqrt{\tilde{q}|F|/r},$$

where the first inequality follows from (5.123) and (5.124), as well as the fact that $\tilde{q} < c_0 r/4$ (so that $\frac{1}{2}\sqrt{c_0} - \sqrt{\tilde{q}/r} > 0$); and the last inequality follows from (5.122). Combining this to the well known inequality that $\sqrt{a} + \sqrt{(b-a)_+} \geq \sqrt{b}$ for any $a, b \geq 0$, we have

$$\sqrt{\omega_0} \geq \sqrt{\omega} - \sqrt{(\omega - \omega_0)_+} \geq \sqrt{\omega} - \big(\frac{1}{2}\sqrt{\omega} - \sqrt{\tilde{q}|F|/r}\big) \geq \frac{1}{2}\sqrt{\omega} + \sqrt{\tilde{q}|F|/r},$$

and (5.121) follows directly. □

5.7.1. *Proof of Lemma 5.9.* Recall that $T(d, F, N) = W'Q^{-1}W - W_N'(Q_{N,N})^{-1}W_N$ where $d = DY$, $W$ and $Q$ are defined in (1.15) which depend on $\mathcal{I} = F \cup N$, and $W_N$ and $Q_{N,N}$ are defined in (1.16). Let $V = S(\beta) \backslash \mathcal{I}$. By definitions,

$$W = Q\beta^{\mathcal{I}} + \xi + u, \quad \text{where } \xi = (B^{\mathcal{I}^{ps},\mathcal{I}})'(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}B^{\mathcal{I}^{ps},V}\beta^V \text{ and } u \sim N(0, Q).$$

Denote $\tilde{W} = Q\beta^{\mathcal{I}} + u$. Introduce a proxy of $T(d, F, N)$ by

$$T_0(d, F, N) = \tilde{W}'Q^{-1}\tilde{W} - (\tilde{W}^N)'(Q^{N,N})^{-1}\tilde{W}^N.$$

Write for short $T = T(d, F, N)$ and $T_0 = T_0(d, F, N)$. To show the claim, it is sufficient to show (a) $|T - T_0| \leq C(\log(p))^{1/\alpha}$ with probability at least $1 - o(1/p)$ and (b) $(T_0|\beta^{\mathcal{I}}) \sim \chi_k^2(\delta_0)$.

Consider (a) first. By direct calculations,

$$|T - T_0| \leq 2\|\xi\| \cdot \big(2\|\beta^{\mathcal{I}}\| + \|Q^{-1}\|\|\xi\| + 2\|Q^{-1/2}\|\|Q^{-1/2}u\|\big). \tag{5.125}$$

First, since $|\mathcal{I}| \leq m$ and $\|\beta\|_\infty \leq a\tau_p \leq C\sqrt{\log(p)}$, $\|\beta^{\mathcal{I}}\| \leq C\sqrt{\log(p)}$. Second, by definitions, $\max\{\|Q^{-1/2}\|, \|Q^{-1}\|\} \leq C$. Last, note that $Q^{-1/2}u \sim N(0, I_{|\mathcal{I}|})$ and so with probability at least $1 - o(1/p)$, $\|Q^{-1/2}u\| \leq C\sqrt{\log(p)}$. Inserting these into (5.125), we have that with probability at least $1 - o(1/p)$,

$$|T - T_0| \leq C\|\xi\|\big(\sqrt{\log(p)} + \|\xi\|\big).$$

We now study $\|\xi\|$. By definitions, it is seen that

$$\|\xi\| \leq \|B^{\mathcal{I}^{ps},\mathcal{I}}\| \cdot \|(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}\| \cdot \|B^{\mathcal{I}^{ps},V}\beta^V\|.$$

First, we have $\|B^{\mathcal{I}^{ps},\mathcal{I}}\| \leq \|B\| \leq C$. Second, since $|\mathcal{I}^{ps}| \leq C$, by RCB, $\lambda_{\min}(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}}) \geq c_1|\mathcal{I}^{ps}|^{-\kappa} \geq C > 0$, and so $\|(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}\| \leq C$. Third, by basic algebra,

$$\|B^{\mathcal{I}^{ps},V}\beta^V\| \leq \sqrt{|\mathcal{I}^{ps}|} \cdot \|B^{\mathcal{I}^{ps},V}\beta^V\|_\infty \leq C\|B^{\mathcal{I}^{ps},V}\|_\infty \cdot \|\beta^V\|_\infty. \tag{5.126}$$

Here, we note that $\|B^{\mathcal{I}^{ps},V}\|_\infty \leq \|B - B^{**}\|_\infty$, where $B^{**}$ is defined in Section 5.1, and where by Lemma 5.1, $\|B - B^{**}\|_\infty \leq C(\log(p))^{-(1-1/\alpha)}$. As a result, $\|B^{\mathcal{I}^{ps},V}\|_\infty \leq C(\log(p))^{-(1-1/\alpha)}$. Inserting this into (5.126) and recalling that $\|\beta^V\|_\infty \leq C\sqrt{\log(p)}$,

$$\|B^{\mathcal{I}^{ps},V}\beta^V\| \leq C(\log(p))^{-(1-1/\alpha)} \cdot \sqrt{\log(p)} = C(\log(p))^{1/\alpha - 1/2}.$$

Combining these gives that $\|\xi\| \leq C(\log(p))^{1/\alpha-1/2}$. This, together with (5.125), implies that

$$|T - T_0| \leq C(\log(p))^{1/\alpha-1/2}[\sqrt{\log(p)} + (\log(p))^{1/\alpha-1/2}] \leq C[(\log(p))^{1/\alpha} + (\log(p))^{2/\alpha-1}],$$

and the claim follows by recalling $\alpha > 1$.

Next, consider (b). Write for short $R = (H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1/2}B^{\mathcal{I}^{ps},\mathcal{I}}$. Also, recall that $F$ and $N$ are subsets of $\mathcal{I}$. We let $R_F$ and $R_N$ be the submatrices of $R$ by restricting the columns to $F$ and $N$, respectively (no restriction on the rows). By definitions, $Q = R'R$ and $u \sim N(0, Q)$, so that we can rewrite $u = R'\tilde{z}$ for some random vector $\tilde{z} \sim N(0, I_{|\mathcal{I}^{ps}|})$. With these notations, we can rewrite $T_0$ as

$$T_0 = (R\beta^{\mathcal{I}} + \tilde{z})'[R(R'R)^{-1}R' - R_N(R_N'R_N)^{-1}R_N'](R\beta^{\mathcal{I}} + \tilde{z}).$$

Therefore, $(T_0|\beta^{\mathcal{I}}) \sim \chi^2_k(\tilde{\delta}_0)$ [21], where $k = \text{rank}(R) - \text{rank}(R_N) \leq |\mathcal{I}|$, and

$$\tilde{\delta}_0 \equiv (R\beta^{\mathcal{I}})'[R(R'R)^{-1}R' - R_N(R_N'R_N)^{-1}R_N'](R\beta^{\mathcal{I}}).$$

By basic algebra, $\tilde{\delta}_0 = \delta_0$. This completes the proof. □

5.8. *Proof of Lemma 2.2.* Viewing $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$, we recall that $\mathcal{I} \lhd \mathcal{U}_p^*$ stands for that $\mathcal{I}$ is a component of $\mathcal{U}_p^*$. The assertion of Lemma 2.2 is that there exists a constant integer $l_0$ such that

$$P\big(|\mathcal{I}| > l_0 \text{ for some } \mathcal{I} \lhd \mathcal{U}_p^*\big) = o(1/p). \tag{5.127}$$

The key to show the claim is the following lemma, which is proved below:

LEMMA 5.10. *There is an event $A_p$ and a constant $C_1 > 0$ such that $P(A_p^c) = o(1/p)$ and that over the event $A_p$, $\|d^{\mathcal{I}^{ps}}\|^2 \geq 5C_1|\mathcal{I}|\log(p)$ for all $\mathcal{I} \lhd \mathcal{U}_p^*$.*

By Lemma 5.10, to show (5.127), it suffices to show

$$P\big(|\mathcal{I}| > l_0 \text{ for some } \mathcal{I} \lhd \mathcal{U}_p^*,\ A_p\big) = o(1/p). \tag{5.128}$$

Now, for each $1 \leq j \leq p$, there is a unique component $\mathcal{I}$ such that $j \in \mathcal{I} \lhd \mathcal{U}_p^*$. Such $\mathcal{I}$ is random, but any of its realization is a connected subgraph of $\mathcal{G}^+$. Therefore,

$$P\big(|\mathcal{I}| > l_0 \text{ for some } \mathcal{I} \lhd \mathcal{U}_p^*,\ A_p\big) \leq \sum_{j=1}^{p} \sum_{l=l_0+1}^{\infty} \sum_{\mathcal{I}: j \in \mathcal{I} \unlhd \mathcal{G}^+, |\mathcal{I}|=l} P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ A_p\big), \tag{5.129}$$

where on the right hand side we have changed the meaning of $\mathcal{I}$ to denote a fixed (non-random) connected subgraph of $\mathcal{G}^+$. We argue that

(a) for each $(j, l)$, the third summation on the right of (5.129) sums over no more than $L_p$ terms;
(b) there are constants $C_2, C_3 > 0$ such that for any $(j, \mathcal{I})$ satisfying $j \in \mathcal{I} \unlhd \mathcal{G}^+$, $P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ A_p\big) \leq L_p\big[p^{-C_2\sqrt{|\mathcal{I}|}} + p^{-C_3|\mathcal{I}|}\big]$.

Once (a) and (b) are proved, then it follows from (5.129) that

$$P\big(|\mathcal{I}| > l_0 \text{ for some } \mathcal{I} \lhd \mathcal{U}_p^*,\ A_p\big) \leq L_p\big[p^{1-C_2\sqrt{l_0}} + p^{1-C_3 l_0}\big],$$

and (5.128) follows by taking $\ell_0$ sufficiently large.

It remains to show (a) and (b). Consider (a) first. Note that the number of connected subgraph $\mathcal{I}$ of size $l$ such that $j \in \mathcal{I} \unlhd \mathcal{G}^+$ is bounded by $C(eK_p^+)^l$

[14], where $K_p^+$ is the maximum degree of $\mathcal{G}^+$. At the same time, by Lemma 5.1 and Lemma 5.2, $K_p^+$ is an $L_p$ term. Combining these gives (a).

Consider (b). Denote $V = \{j : B^{**}(i,j) \neq 0, \text{ for some } i \in \mathcal{I}^{ps}\}$, where $B^{**}$ is defined in Section 5.1. Write for short $d_1 = d^{\mathcal{I}^{ps}}$, $B_1 = B^{\mathcal{I}^{ps},V}$ and $H_1 = H^{\mathcal{I}^{ps},\mathcal{I}^{ps}}$. With these notations and by Lemma 5.10, (b) reduces to

$$P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \|d_1\|^2 \geq 5C_1|\mathcal{I}|\log(p)\big) \leq L_p\big[p^{-C_2\sqrt{|\mathcal{I}|}} + p^{-C_3|\mathcal{I}|}\big]. \tag{5.130}$$

We now show (5.130). Note that $d_1 = B_1\beta^V + \xi + \tilde{z}$, where $\xi = [(B - B^{**})\beta]^{\mathcal{I}^{ps}}$ and $\tilde{z} \sim N(0, H_1)$. For preparation, we claim that

$$\|\xi\|^2 = |\mathcal{I}| \cdot o(\log(p)). \tag{5.131}$$

In fact, first since $\ell^{ps}$ is finite, $|\mathcal{I}^{ps}| \leq C|\mathcal{I}|$ and it follows that $\|\xi\|^2 \leq C|\mathcal{I}| \cdot \|\xi\|_\infty^2$. Second, by Lemma 5.1, $\|B - B^{**}\|_\infty = o(1)$. Since $\|\beta\|_\infty \leq a\tau_p \leq C\sqrt{\log(p)}$, it follows that $\|\xi\|_\infty \leq \|B - B^{**}\|_\infty\|\beta\|_\infty = o(\sqrt{\log(p)})$. Combining these gives (5.131).

Now, combining (5.130) and (5.131) and using the well-known inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$, we find that for sufficiently large $p$,

$$\begin{aligned} &P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \|d_1\|^2 \geq 5C_1|\mathcal{I}|\log(p)\big) \\ \leq\ & P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \|B_1\beta^V + \tilde{z}\|^2 \geq 4C_1|\mathcal{I}|\log(p)\big) \\ \leq\ & P\big(j \in \mathcal{I} \lhd \mathcal{U}_p^*,\ \|B_1\beta^V\|^2 + \|\tilde{z}\|^2 \geq 2C_1|\mathcal{I}|\log(p)\big) \\ \leq\ & P\big(\|B_1\beta^V\|^2 \geq C_1|\mathcal{I}|\log(p)\big) + P\big(\|\tilde{z}\|^2 \geq C_1|\mathcal{I}|\log(p)\big) \equiv I + II. \end{aligned} \tag{5.132}$$

We now analyze $I$ and $II$ separately. Consider $I$ first. We claim there is a constant $C_4 > 0$, not depending on $|\mathcal{I}|$, such that $\|B_1\beta^V\| \leq \sqrt{C_4\log(p)}\|\beta^V\|_0$. To see this, note that $\|B_1\beta^V\| \leq \|B_1\beta^V\|_1 \leq \|B_1\|_1\|\beta^V\|_1$, where $\|B_1\|_1 \leq \|B\|_1 \leq C$, with $C > 0$ a constant independent of $|\mathcal{I}|$. At the same time, $\|\beta^V\|_1 \leq a\tau_p\|\beta^V\|_0$. So the argument holds for $C_4 = 2ra^2C^2$. Additionally, $\|\beta^V\|_0$ has a multinomial distribution, where the number of trials is $|V| \leq L_p$ and the success probability is $\epsilon_p = p^{-\vartheta}$. Combining these, we have

$$I \leq P\big(\|\beta^V\|_0 \geq \sqrt{(C_1/C_4)|\mathcal{I}|}\big) \leq L_p p^{-\vartheta\lfloor\sqrt{(C_1/C_4)|\mathcal{I}|}\rfloor}, \tag{5.133}$$

where $\lfloor x \rfloor$ denotes the the largest integer $k$ such that $k \leq x$.

Next, consider $II$. Note that $\|H_1\| \leq \|H\| \leq C_5$, where $C_5 > 0$ is a constant independent of $|\mathcal{I}|$. It follows that $\|\tilde{z}\|^2 \leq C_5^{-1}\|H_1^{-1/2}\tilde{z}\|^2$, where $\|H_1^{-1/2}\tilde{z}\|^2$ has a $\chi^2$-distribution with degree of freedom $|\mathcal{I}^{ps}| \leq C|\mathcal{I}|$. Using the property of $\chi^2$-distributions,

$$II \leq P\big(\|H_1^{-1/2}\tilde{z}\|^2 \geq C_1C_5|\mathcal{I}|\log(p)\big) \leq L_p p^{-(C_1C_5/2)|\mathcal{I}|}. \tag{5.134}$$

Inserting (5.133) and (5.134) into (5.132), (5.130) follows by taking $C_2 > \vartheta\sqrt{C_1/C_4}$ and $C_3 > C_1C_5/2$. □

5.8.1. *Proof of Lemma 5.10.* For preparation, we need some notations. First, for a constant $\delta_0 > 0$ to be determinded, define the $p \times p$ matrices $\tilde{B}$ and $\tilde{H}$ by

$$\tilde{B}(i,j) = B(i,j)1\{|B(i,j)| > \delta_0\}, \qquad \tilde{H}(i,j) = H(i,j)1\{|H(i,j)| > \delta_0\}.$$

Second, view $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$. Note that in the $PS$-step, each $\mathcal{G}_t$ is a connected subgraph of $\mathcal{G}^+$. Hence, any $\mathcal{G}_t$ that passed the test must be contained as a whole in one component of $\mathcal{U}_p^*$. It follows that for any $\mathcal{I} \lhd \mathcal{U}_p^*$, there exists a (random) set $\mathcal{T} \subset \{1, \cdots, T\}$ such that $\mathcal{I} = \cup_{t\in\mathcal{T}}\mathcal{G}_t$. Therefore, we write

$$\mathcal{I} = \cup_{i=1}^{\hat{s}_0} V_i,$$

where each $V_i = \mathcal{G}_t$ for some $t \in \mathcal{T}$, and these $V_i$'s are listed in the order they were tested. Denote $\hat{N}_i = \mathcal{U}^{(t-1)} \cap \mathcal{G}_t$ and $\hat{F}_i = \mathcal{G}_t \backslash \hat{N}_i$. Let $W_{(i)}$ and $Q_{(i)}$ be the vector $W$ and matrix $Q$ in (1.15). From basic algebra, the test statistic can be rewritten as

$$T(d, \hat{F}_i, \hat{N}_i) = \|u_{(i)}\|^2, \qquad u_{(i)} \equiv \Sigma_{(i)}^{-1/2}[W_{(i)}^{\hat{F}_i} - Q_{(i)}^{\hat{F}_i,\hat{N}_i}(Q_{(i)}^{\hat{N}_i,\hat{N}_i})^{-1}W_{(i)}^{\hat{N}_i}], \tag{5.135}$$

where $\Sigma_{(i)} = Q_{(i)}^{\hat{F}_i,\hat{F}_i} - Q_{(i)}^{\hat{F}_i,\hat{N}_i}[Q_{(i)}^{\hat{N}_i,\hat{N}_i}]^{-1}Q_{(i)}^{\hat{N}_i,\hat{F}_i}$.

Third, define

$$W_{(i)}^* = (\tilde{B}^{V_i^{ps},V_i})'(\tilde{H}^{V_i^{ps},V_i^{ps}})^{-1}d^{V_i^{ps}},$$

and $u_{(i)}^*$ as in (5.135) with $W_{(i)}$ replaced by $W_{(i)}^*$. Let $u$ be the $|\mathcal{I}| \times 1$ vector by putting $\{u_{(i)}, 1 \le i \le \hat{s}_0\}$ together, and define $u^*$ similarly.

With these notations, to show the claim, it suffices to show there exist positive constants $C_6, C_7$ such that with probability at least $1 - o(1/p)$, for any $\mathcal{I} \lhd \mathcal{U}_p^*$,

$$\|u^*\|^2 \ge C_6|\mathcal{I}|\log(p), \tag{5.136}$$

and

$$\|u^*\|^2 \le C_7\|d^{\mathcal{I}^{ps}}\|^2. \tag{5.137}$$

Consider (5.136) first. Since each $V_i$ passed the test, $\|u_{(i)}\|^2 \ge t(\hat{F}_i, \hat{N}_i)$. If $t(\hat{F}_i, \hat{N}_i)$ is chosen from (1.31), $t(\hat{F}_i, \hat{N}_i) \ge 2q_0\log(p) \ge 2(q_0/m)|\hat{F}_i|\log(p)$;

otherwise it is chosen from (1.36), then $t(\hat{F}_i, \hat{N}_i) \geq 2\tilde{q}|\hat{F}_i|\log(p)$. In both cases, there is a constant $q > 0$ such that

$$\|u_{(i)}\|^2 \geq 2q|\hat{F}_i|\log(p), \qquad 1 \leq i \leq \hat{s}_0.$$

In addition, it is easy to see that $\cup_i \hat{F}_i$ is a partition of $\mathcal{I}$. It follows that

$$\|u\|^2 = \sum_{i=1}^{\hat{s}_0} \|u_{(i)}\|^2 \geq 2q|\mathcal{I}|\log(p). \tag{5.138}$$

At the same time, let $A_p$ be the event $\{\|d\|_\infty \leq C_0\sqrt{\log(p)}\}$, where we argue that when $C_0$ is sufficiently large, $P(A_p^c) = o(1/p)$. To see this, recall that $d = B\beta + H^{1/2}\tilde{z}$, where $\tilde{z} \sim N(0, I_p)$. By the assumptions, $\|B\|_\infty \leq C$, $\|\beta\|_\infty \leq C\sqrt{\log(p)}$ and $\|H\|_\infty \leq C$. Therefore, $\|d\|_\infty \leq C(\sqrt{\log(p)} + \|\tilde{z}\|_\infty)$. It is well-known that $P(\|\tilde{z}\|_\infty > \sqrt{2a\log(p)}) = L_p p^{-a}$ for any $a > 0$. Hence, when $C_0$ is sufficiently large, $P(A_p^c) = o(1/p)$.

We shall show that over the event $A_p$, by choosing $\delta_0$ a sufficiently small constant,

$$\|u - u^*\|^2 \leq q|\mathcal{I}|\log(p)/2. \tag{5.139}$$

Once this is proved, combining (5.138) and (5.139), and applying the inequality $(a+b)^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$, we have

$$2q|\mathcal{I}|\log(p) \leq \|u\|^2 \leq 2(\|u^*\|^2 + \|u - u^*\|^2) \leq 2\|u^*\|^2 + q|\mathcal{I}|\log(p).$$

Hence, (5.136) holds with $C_6 = q/2$.

What remains is to prove (5.139). It follows from $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$ and $|V_i| \leq m \leq g$ that $\|(G^{V_i,V_i})^{-1}\| \leq c_0^{-1}$. As a result, $\|Q_{(i)}^{-1}\| \leq C$. Also, $\Sigma_{(i)}^{-1}$ is a submatrix of $Q_{(i)}^{-1}$; and hence $\|\Sigma_{(i)}^{-1}\| \leq C$. This implies

$$\|u_{(i)} - u_{(i)}^*\| \leq C\|W_{(i)} - W_{(i)}^*\|, \qquad 1 \leq i \leq \hat{s}_0. \tag{5.140}$$

Since $B$ enjoys a polynomial off-diagonal decay with rate $\alpha$, $\|(B - \tilde{B})^{V_i^{ps},V_i}\|_\infty \leq C\delta_0^{1-1/\alpha}$. Noting that $|V_i^{ps}| \leq C$, this implies $\|(B - \tilde{B})^{V_i^{ps},V_i}\| \leq C\delta_0^{1-1/\alpha}$. Similarly, we can derive $\|(H - \tilde{H})^{V_i^{ps},V_i^{ps}}\| \leq C\delta_0^{1-1/\alpha}$. These together imply

$$\|W_{(i)} - W_{(i)}^*\| \leq C\delta_0^{1-1/\alpha}\|d^{V_i^{ps}}\| \leq C\delta_0^{1-1/\alpha}\|d\|_\infty, \quad 1 \leq i \leq \hat{s}_0, \tag{5.141}$$

where in the last inequality we use the facts that $|V_i^{ps}| \leq C$ and $\|d^{V_i^{ps}}\|_\infty \leq \|d\|_\infty$. Combining (5.140) and (5.141), over the event $A_p$,

$$\|u_{(i)} - u_{(i)}^*\|^2 \leq C\delta_0^{2(1-1/\alpha)}\log(p), \qquad 1 \leq i \leq \hat{s}_0.$$

Noting that $\alpha > 1$, we can choose a sufficiently small $\delta_0$ such that $C\delta_0^{2(1-1/\alpha)} \leq q/2$, and (5.139) follows by noting $|\hat{s}_0| \leq |\mathcal{I}|$.

Next, consider (5.137). We write

$$u^* = \Xi\Gamma\Theta\ d^{\mathcal{I}^{ps}},$$

where the matrices $\Xi$, $\Gamma$ and $\Theta$ are defined as follows: $\Xi$ is a block-wise diagonal matrix with the $i$-th block equals to $\Sigma_{(i)}^{-1}$. $\Gamma$ is a $|\mathcal{I}| \times (\sum_{i=1}^{\hat{s}_0} |V_i|)$ matrix, with the $(\hat{F}_i, V_i)$-block is given by

$$\Gamma^{\hat{F}_i, V_i} = \left[I,\ -Q_{(i)}^{\hat{F}_i, \hat{N}_i}(Q_{(i)}^{\hat{N}_i, \hat{N}_i})^{-1}\right].$$

and 0 elsewhere. $\Theta$ is a $(\sum_{i=1}^{\hat{s}_0} |V_i|) \times |\mathcal{I}^{ps}|$ matrix, with the $(V_i, V_i^{ps})$-block

$$\Theta^{V_i, V_i^{ps}} = (\tilde{B}^{V_i^{ps}, V_i})'(\tilde{H}^{V_i^{ps}, V_i^{ps}})^{-1},$$

and 0 elsewhere.

Note that these matrices are random (they depend on $\mathcal{U}_p^*$ and $\mathcal{I}$). Below, we show that for any realization of $\mathcal{U}_p^*$ and any component $\mathcal{I} \lhd \mathcal{U}_p^*$,

$$\|\Xi\Gamma\Theta\| \leq C. \tag{5.142}$$

Once (5.142) is proved, (5.137) follows by letting $C_7 = C^2$.

We now show (5.142). Since $\|\Xi\Gamma\Theta\| \leq \|\Xi\|\|\Gamma\|\|\Theta\|$, it suffices to show

$$\|\Xi\|, \|\Gamma\|, \|\Theta\| \leq C.$$

First, $\|\Xi\| \leq \max_i \|Q_{(i)}^{-1}\| \leq C$. Second, the entries in $\Gamma$ and $\Theta$ have a uniform upper bound in magnitude, and each row and column of $\Gamma$ has $\leq m$ non-zero entries. So $\|\Gamma\| \leq C$. Finally, each row of $\Theta$ has no more than $2m\ell^{ps}$ entries; as a result, to show $\|\Theta\| \leq C$, we only need to prove that each column of $\Theta$ also has a bounded number of non-zero entries.

Towards this end, write for short $\tilde{B}_{(i)} = \tilde{B}^{V_i^{ps}, V_i}$ and $\tilde{H}_{(i)} = \tilde{H}^{V_i^{ps}, V_i^{ps}}$ for each $1 \leq i \leq \hat{s}_0$. By definition,

$$\Theta(k, j) = \sum_{j' \in V_i^{ps}} \tilde{B}_{(i)}(j', k)\tilde{H}_{(i)}^{-1}(j', j), \qquad k \in V_i,\ j \in V_i^{ps}.$$

First, given the chosen $\delta_0$, each row or column of $\tilde{B}$ and $\tilde{H}$ has $\leq L_0$ non-zero entries, where $L_0$ is a constant integer. Therefore, for each $j'$, the number of $k$ such that $\tilde{B}(j', k) \neq 0$ is upper bounded by $L_0$. Second, we define a graph $\mathcal{G} = \mathcal{G}(\delta_0)$ where there is an edge between nodes $j$ and $j'$ if and only if

$\tilde{H}(j,j') \neq 0$. For each $1 \leq i \leq \hat{s}_0$, let $\mathcal{G}_i$ be the restriction of $\mathcal{G}$ to the nodes in $V_i^{ps}$. We see that $\tilde{H}_{(i)}$ is block-diagonal with each block corresponding to a component of $\mathcal{G}_i$, and so is $(\tilde{H}_{(i)})^{-1}$. This means $(\tilde{H}_{(i)})^{-1}(j',j)$ can be non-zero only when $j$ and $j'$ belong to the same component of $\mathcal{G}_i$. Since $|V_i^{ps}| \leq 2m\ell^{ps}$ for all $i$, necessarily, there exits a path in $\mathcal{G}$ of length $\leq 2m\ell^{ps}$ that connects $j$ and $j'$. Third, since $\mathcal{G}$ is $L_0$-sparse, for each $j$, the number of $j'$ that is connected to $j$ with a path of length $\leq 2m\ell^{ps}$ is upper bounded by $L_0^{2m\ell^{ps}}$. In summary, for each fixed $j$, there are no more than $L_0 \cdot L_0^{2m\ell^{ps}}$ nodes $k$ such that $\Theta(k,j) \neq 0$, i.e., each column of $\Theta$ has $\leq L_0^{2m\ell^{ps}+1}$ nonzero entries and the claim follows. $\square$

5.9. *Proof of Lemma 2.3.* Fix $\mathcal{I}$ and recall that $\mathcal{J} = \{j : D(i,j) \neq 0, \text{ for some } i \in \mathcal{I}\}$. In this lemma, $\mathcal{I}^{pe}$ is as in Definition 1.6, but $\mathcal{J}^{pe}$ is redefined as $\mathcal{J}^{pe} = \{j : D(i,j) \neq 0, \text{ for some } i \in \mathcal{I}^{pe}\}$. Denote $M = |\mathcal{J}^{pe}| - |\mathcal{I}^{pe}|$ and write $G^{\mathcal{J}^{pe},\mathcal{J}^{pe}} = G_1$ for short. Let $\mathcal{F}$ be the mapping from $\mathcal{J}^{pe}$ to $\{1,\cdots,|\mathcal{J}^{pe}|\}$ that maps each $j \in \mathcal{J}^{pe}$ to its order in $\mathcal{J}^{pe}$. Denote $\mathcal{I}_1 = \mathcal{F}(\mathcal{I})$. By these notations, the claim reduces to: for any $|\mathcal{J}^{pe}| \times M$ matrix $U$ whose columns contain an orthonormal basis of $Null(\mathcal{I}^{pe},\mathcal{J}^{pe})$,

$$\left\|[U(U'G_1^{-1}U)^{-1}U']^{\mathcal{I}_1,\mathcal{I}_1}\right\| = o(1).$$

It suffices to show

$$\left\|[U(U'G_1^{-1}U)^{-1}U']^{\mathcal{I}_1,\mathcal{I}_1}\right\| \leq C(\ell^{pe})^{-\gamma}, \tag{5.143}$$

where $\gamma > 0$ is the same as in $\mathcal{M}_p^*(\gamma,g,c_0,A_1)$. In fact, once this is proved, the claim follows by noting that $\ell^{pe} = (\log(p))^{\nu} \to \infty$.

We now show (5.143). By elementary algebra,

$$\left\|[U(U'G_1^{-1}U)^{-1}U']^{\mathcal{I}_1,\mathcal{I}_1}\right\| \leq \left\|(U'G_1^{-1}U)^{-1}\right\| \left\|(UU')^{\mathcal{I}_1,\mathcal{I}_1}\right\|. \tag{5.144}$$

Consider $\left\|(U'G_1^{-1}U)^{-1}\right\|$ first. Since $U'U$ is an identity matrix, we have $\|(U'G_1^{-1}U)^{-1}\| = [\lambda_{\min}(U'G_1^{-1}U)]^{-1} \leq [\lambda_{\min}(G_1^{-1})]^{-1} = \|G_1\|$. Additionally, the assumption $G \in \mathcal{M}_p^*(\gamma,g,c_0,A_1)$ implies that $\|G_1\| \leq A_1 \sum_{j=1}^{|\mathcal{J}^{pe}|} j^{-\gamma} \leq C|\mathcal{J}^{pe}|^{1-\gamma}$. Last, when $|\mathcal{I}| \leq l_0$, $2\ell^{pe}+1 \leq |\mathcal{J}^{pe}| \leq (2\ell^{pe}+1)l_0$. Combining the above yields

$$\left\|(U'G_1^{-1}U)^{-1}\right\| \leq C(\ell^{pe})^{1-\gamma}. \tag{5.145}$$

Next, consider $\left\|(UU')^{\mathcal{I}_1,\mathcal{I}_1}\right\|$. Note that $\|(UU')^{\mathcal{I}_1,\mathcal{I}_1}\| \leq |\mathcal{I}_1| \cdot \max_{i,i'\in\mathcal{I}_1} |(UU')(i,i')|$, where $\max_{i,i'\in\mathcal{I}_1} |U'U(i,i')| \leq M \cdot \max_{i\in\mathcal{I}_1, 1\leq j\leq M} |U(i,j)|^2$. Here $|\mathcal{I}_1| = |\mathcal{I}| \leq l_0$ and $M \leq h|\mathcal{I}| \leq hl_0$. It follows that

$$\left\|(UU')^{\mathcal{I}_1,\mathcal{I}_1}\right\| \leq C \max_{i\in\mathcal{I}_1, 1\leq j\leq M} |U(i,j)|^2. \tag{5.146}$$

The following lemma is proved in Appendix A.

LEMMA 5.11. *Under the conditioins of Lemma 2.3, for any $\mathcal{I} \trianglelefteq \mathcal{G}^+$ such that $|\mathcal{I}| \leq l_0$, and any matrix $U$ whose columns form an orthonormal basis of $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$,*

$$\max_{i \in \mathcal{F}(\mathcal{I}), 1 \leq j \leq |\mathcal{J}^{pe}| - |\mathcal{I}^{pe}|} |U(i,j)|^2 \leq C(\ell^{pe})^{-1}.$$

Using Lemma 5.11, it follows from (5.146) that

$$\left\| (UU')^{\mathcal{I}_1, \mathcal{I}_1} \right\| \leq C(\ell^{pe})^{-1}. \tag{5.147}$$

Inserting (5.145) and (5.147) into (5.144), we obtain (5.143). □

5.10. *Proof of Lemma 2.4.* Write for short

$$M_1 = \sum_{j=1}^{p} \sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0} P\big(\mathcal{I} \triangleleft \mathcal{U}_p^*,\ A_p \cap E_{p,\mathcal{I}}^c\big), \quad M_2 = \sum_{k=1}^{p} P\big(\beta_k \neq 0,\ k \notin \mathcal{U}_p^*\big).$$

With these notations, the claim reduces to $M_1 \leq L_p \cdot M_2$.

The key is to prove

(a) for each $\mathcal{I} \trianglelefteq \mathcal{G}^+$, over the event $\{\mathcal{I} \triangleleft \mathcal{U}_p^*, A_p \cap E_{p,\mathcal{I}}^c\}$, it always holds that $(S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe})) \setminus \mathcal{U}_p^* \neq \emptyset$;
(b) for each $k$, there are no more than $L_p$ different $\mathcal{I}$ such that $\mathcal{I} \trianglelefteq \mathcal{G}^+$, $|\mathcal{I}| \leq l_0$ and $k \in \mathcal{E}(\mathcal{I}^{pe})$.

Once (a) and (b) are proved, the claim follows easily. To see the point, we note that

$$P\left((S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe})) \setminus \mathcal{U}_p^* \neq \emptyset\right) \leq \sum_{k \in \mathcal{E}(\mathcal{I}^{pe})} P\big(\beta_k \neq 0,\ k \notin \mathcal{U}_p^*\big).$$

Combining this with (a), we have

$$M_1 \leq \sum_{j=1}^{p} \sum_{\mathcal{I}: j \in \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0} \sum_{k \in \mathcal{E}(\mathcal{I}^{pe})} P\big(\beta_k \neq 0,\ k \notin \mathcal{U}_p^*\big).$$

By re-organizing the summation, the right hand side is equal to

$$\sum_{k=1}^{p} \sum_{\mathcal{I}: \mathcal{I} \trianglelefteq \mathcal{G}^+, |\mathcal{I}| \leq l_0, k \in \mathcal{E}(\mathcal{I}^{pe})} |\mathcal{I}| \cdot P\big(\beta_k \neq 0,\ k \notin \mathcal{U}_p^*\big),$$

which $\leq L_p \cdot M_2$ by (b), and the claim follows.

We now show (a) and (b). Consider (a) first. Fix $\mathcal{I} \trianglelefteq \mathcal{G}^+$. Suppose (a) does not hold, i.e., the following event

$$\big\{\mathcal{I} \lhd \mathcal{U}_p^*,\ (S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe}))\backslash \mathcal{U}_p^* = \emptyset,\ A_p \cap E_{p,\mathcal{I}}^c\big\}$$

is non-empty. View $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$. Applying Lemma 5.2 to $V = \mathcal{U}_p^*$, we find that $\mathcal{I} \lhd \mathcal{U}_p^*$ implies $(\mathcal{U}_p^* \backslash \mathcal{I}) \cap \mathcal{E}(\mathcal{I}^{pe}) = \emptyset$. Therefore, the following event

$$\big\{(\mathcal{U}_p^* \backslash \mathcal{I}) \cap \mathcal{E}(\mathcal{I}^{pe}) = \emptyset,\ (S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe}))\backslash \mathcal{U}_p^* = \emptyset,\ A_p \cap E_{p,\mathcal{I}}^c\big\} \tag{5.148}$$

is non-empty. Note that $\mathcal{I} \subset \mathcal{E}(\mathcal{I}^{pe})$. From basic set operations, $(\mathcal{U}_p^* \backslash \mathcal{I}) \cap \mathcal{E}(\mathcal{I}^{pe}) = \emptyset$ and $(S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe}))\backslash \mathcal{U}_p^* = \emptyset$ together imply

$$\big(S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe})\big) \subset \mathcal{I}.$$

By definition, this belongs to the event $E_{p,\mathcal{I}}$. Hence, the event in (5.148) is empty, which is a contradiction.

Consider (b) next. Fix $k$ and denote $\mathcal{K}$ the collection of $\mathcal{I}$ satisfying the conditions in (b). Let $V = \big\{1 \leq i \leq p:\ \ k \in \mathcal{E}(\{i\}^{pe})\big\}$. Since $\mathcal{E}(\mathcal{I}^{pe}) = \cup_{i \in \mathcal{I}} \mathcal{E}(\{i\}^{pe})$, we observe that

$$\mathcal{K} = \cup_{i \in V} \mathcal{K}_i, \qquad \text{where} \quad \mathcal{K}_i \equiv \big\{\mathcal{I}:\ \mathcal{I} \trianglelefteq \mathcal{G}^+,\ |\mathcal{I}| \leq l_0,\ i \in \mathcal{I}\big\}.$$

Note that by Lemma 5.1 and 5.2, $\mathcal{G}^*$ is $K_p$-sparse and $\mathcal{G}^+$ is $K_p^+$-sparse, where both $K_p$ and $K_p^+$ are $L_p$ terms. First, we bound $|V|$: By definition, $k \in \mathcal{E}(\{i\}^{pe})$ if and only if there exits a node $k' \in \{i\}^{pe}$ such that $k'$ and $k$ are connected by a length-1 path in $\mathcal{G}^*$. Since $\mathcal{G}^*$ is $K_p$-sparse, given $k$, the number of such $k'$ is bounded by $K_p$. In addition, for each $k'$, there are no more than $(2\ell^{pe}+1)$ nodes $i$ such that $k' \in \{i\}^{pe}$. Hence, $|V| \leq (2\ell^{pe}+1)K_p$. Second, we bound $\max_{i \in V} |\mathcal{K}_i|$: For each node $i \in V$, there are no more than $C(eK_p^+)^{l_0}$ connected subgraph of $\mathcal{G}^+$ that contain $i$ and have a size $\leq l_0$ [14], i.e., $|\mathcal{K}_i| \leq C(eK_p^+)^{l_0}$. Combining the two parts, $|\mathcal{K}| \leq K_p(2\ell^{pe}+1) \cdot C(eK_p^+)^{l_0}$, which is an $L_p$ term. □

5.11. *Proof of Lemma 2.5.* Let $V_1 = S(\beta) \cap \mathcal{E}(\mathcal{I}^{pe})$ and $V_2 = S(\beta) \backslash \mathcal{E}(\mathcal{I}^{pe})$. We have $(B\beta)^{\mathcal{I}^{pe}} = B^{\mathcal{I}^{pe},V_1}\beta^{V_1} + \zeta$, where $\zeta = B^{\mathcal{I}^{pe},V_2}\beta^{V_2}$. Note that over the event $E_{p,\mathcal{I}}$, $V_1 \subset \mathcal{I}$. It follows that $B^{\mathcal{I}^{pe},V_1}\beta^{V_1} = B^{\mathcal{I}^{pe},\mathcal{I}}\beta^{\mathcal{I}}$. Combining these, to show the claim, it is sufficient to show

$$\|\zeta\| \leq C(\ell^{pe})^{1/2}[\log(p)]^{-(1-1/\alpha)}\tau_p. \tag{5.149}$$

Recall the matrix $B^{**}$ defined in Section 5.1. Since $B^{**}(i,j)=0$ for $j\in V_2$, we have $\|B^{\mathcal{I}^{pe},V_2}\|_\infty \leq \|B-B^{**}\|_\infty$, where by Lemma 5.1, $\|B-B^{**}\|_\infty \leq C[\log(p)]^{-(1-1/\alpha)}$. Moreover, $\|\beta\|_\infty \leq a\tau_p$. Consequently,

$$\|\zeta\|_\infty \leq \|B-B^{**}\|_\infty \|\beta^{V_2}\|_\infty \leq C[\log(p)]^{-(1-1/\alpha)}\tau_p. \tag{5.150}$$

At the same time, note that $|\mathcal{I}^{pe}| \leq l_0(2\ell^{pe}+1) \leq C\ell^{pe}$. It follows from the Cauchy-Schwartz inequality that $\|\zeta\| \leq \sqrt{|\mathcal{I}^{pe}|}\|\zeta\|_\infty \leq C(\ell^{pe})^{1/2}\|\zeta\|_\infty$. Combining this with (5.150) gives the claim. □

5.12. *Proof of Lemma 2.6.* Fix $(j,V_0,V_1,\mathcal{I})$ and write for short $\rho_j(V_0,V_1)=\rho_j(V_0,V_1;\mathcal{I})$ and $\rho_j^*=\rho_j^*(\vartheta,r,G)$. The goal is to show $\rho_j(V_0,V_1)\geq \rho_j^*+o(1)$. We show this for the case $V_0\neq V_1$ and the case $V_0=V_1$ separately.

Consider the first case. By definition, $\rho_j^*\leq \rho(V_0,V_1)$, where $\rho(V_0,V_1)$ is as in (1.26). Therefore, it suffices to show

$$\rho_j(V_0,V_1)=\rho(V_0,V_1)+o(1). \tag{5.151}$$

Introduce the function

$$f(x)=\max\{|V_0|,|V_1|\}\vartheta+\frac{1}{4}\big[\big(\sqrt{x}-\big||V_0|-|V_1|\big|\vartheta/\sqrt{x}\big)_+\big]^2, \qquad x>0.$$

Then $\rho_j(V_0,V_1)=f(\varpi_j r)$ and $\rho(V_0,V_1)=f(\varpi^* r)$, where $\varpi_j=\varpi_j(V_0,V_1;\mathcal{I})$ and $\varpi^*=\varpi^*(V_0,V_1)$, defined in (2.59) and (1.25) respectively. Since $f(x)$ is an increasing function and $|f(x)-f(y)|\leq |x-y|/4$ for all $x,y>0$, to show (5.151), it suffices to show

$$\varpi_j \geq \varpi^*+o(1). \tag{5.152}$$

Now, we show (5.152). Introduce the quantity $\varpi=\min_{j\in(V_0\cup V_1)}\varpi_j$. Write $B_1=B^{\mathcal{I}^{ps},\mathcal{I}}$, $H_1=H^{\mathcal{I}^{ps},\mathcal{I}^{ps}}$ and $Q_1=B_1'H_1^{-1}B_1$. Given any $C>0$, define $\Theta(C)$ as the collection of vectors $\xi\in\mathbb{R}^{|\mathcal{I}|}$ such that for all $i$, either $\xi_i^{(k)}=0$ or $|\xi_i^{(k)}|\geq 1$, and that $\mathrm{Supp}(\xi^{(k)})=V_k$, $\|\xi^{(k)}\|_\infty\leq C$, for $k=0,1$. Denote $\Theta=\Theta(\infty)$. By these notations and the definitions of $\varpi_j$ and $\varpi^*$, we have

$$\varpi=\min_{(\xi^{(0)},\xi^{(1)}):\ \xi^{(k)}\in\Theta,k=0,1;\mathrm{sgn}(\xi^{(0)})\neq\mathrm{sgn}(\xi^{(1)})}(\xi^{(1)}-\xi^{(0)})'Q_1(\xi^{(1)}-\xi^{(0)}),$$
$$\varpi^*=\min_{(\xi^{(0)},\xi^{(1)}):\ \xi^{(k)}\in\Theta(a),k=0,1;\mathrm{sgn}(\xi^{(0)})\neq\mathrm{sgn}(\xi^{(1)})}(\xi^{(1)}-\xi^{(0)})'G^{\mathcal{I},\mathcal{I}}(\xi^{(1)}-\xi^{(0)}).$$

First, since $a>a_g^*(G)$, in the expression of $\varpi^*$, $\Theta(a)$ can be replaced by $\Theta(C)$ for any $C\geq a$. Second, since $\lambda_{\min}(Q_1)\geq C$, from basic properties of

the quadratic programming, there exists a constant $a_0 > 0$ such that for any $(\xi_*^{(0)}, \xi_*^{(1)})$, a minimizer in the expression of $\varpi$, $\max\{\|\xi_*^{(0)}\|_\infty, \|\xi_*^{(1)}\|_\infty\} \leq a_0$. Therefore, in the expression of $\varpi$, $\Theta$ can be replaced by $\Theta(C)$ for any $C \geq a_0$. Now, let $a_1 = \max\{a_0, a\}$ and we can unify the constraints in two expressions to that $\xi^{(k)} \in \Theta(a_1)$, for $k = 0, 1$, and $\mathrm{sgn}(\xi^{(0)}) \neq \mathrm{sgn}(\xi^{(1)})$. It follows that

$$|\varpi - \varpi^*| \leq \max_{\xi \in \mathbb{R}^{|\mathcal{I}|}: \|\xi\|_\infty \leq 2a_1} \left|\xi'(G^{\mathcal{I},\mathcal{I}} - Q_1)\xi\right| \leq C\|G^{\mathcal{I},\mathcal{I}} - Q_1\|. \tag{5.153}$$

Note that $Q_1$ is the Fisher Information Matrix associated with model $d_1 \sim N(B_1\beta^{\mathcal{I}}, H_1)$, by Lemma 1.2 and Lemma 2.3, $\|G^{\mathcal{I},\mathcal{I}} - Q_1\| = o(1)$. Plugging this into (5.153) gives $|\varpi - \varpi^*| = o(1)$. Hence, $\varpi_j \geq \varpi \geq \varpi^* + o(1)$ and (5.152) follows.

Next, consider the case $V_0 = V_1$. Pick an arbitrary minimizer in the definition of $\varpi_j$, denoted as $(\xi_*^{(0)}, \xi_*^{(1)})$, and define $F = \{k : \mathrm{sgn}(\xi_{*k}^{(0)}) \neq \mathrm{sgn}(\xi_{*k}^{(1)})\}$ and $N = V_0 \backslash F$. It is seen that $j \in F$. By Lemma 5.3, $\rho_j^* \leq \psi(F, N)$, where $\psi(F, N)$ is defined in (1.33). Hence, it suffices to show

$$\rho_j(V_0, V_1) \geq \psi(F, N) + o(1). \tag{5.154}$$

On one hand, when $|V_0| = |V_1|$, the function $f$ introduced above is equal to $|V_0|\vartheta + x/4$ and hence

$$\rho_j(V_0, V_1) = f(\varpi_j r) = |V_0|\vartheta + \varpi_j r/4.$$

On the other hand, using the expression of $\psi(F, N)$ in (1.33) and noting that $|F| \geq 1$,

$$\psi(F, N) \leq (|F| + |N|)\vartheta + \omega r/4 = |V_0|\vartheta + \omega r/4,$$

where $\omega = \omega(F, N)$ is defined in (1.34). Therefore, to show (5.154), it suffices to show

$$\varpi_j \geq \omega + o(1). \tag{5.155}$$

Now, we show (5.155). From the definition (1.34) and basic algebra, we can write

$$\omega = \min_{\xi \in \mathbb{R}^{|\mathcal{I}|}:\ \xi_i = 0, i \notin V_0; |\xi_i| \geq 1, i \in F} \xi' G^{\mathcal{I},\mathcal{I}} \xi.$$

Denote $\xi_* = \xi_*^{(1)} - \xi_*^{(0)}$. By our construction, $\varpi_j = \xi_*' Q_1 \xi_*$, $\xi_{*i} = 0$ for $i \notin V_0$, and $|\xi_{*i}| \geq 2$ for $i \in F$. As a result,

$$\xi_*' G^{\mathcal{I},\mathcal{I}} \xi_* \geq \omega. \tag{5.156}$$

At the same time, we have seen in the derivation of (5.153) that there exists a constant $a_0 > 0$ such that $\|\xi_*^{(0)}\|_\infty, \|\xi_*^{(1)}\|_\infty \le a_0$ and $\|G^{\mathcal{I},\mathcal{I}} - Q_1\| = o(1)$. Therefore, $\|\xi_*\|^2 \le 2a_0|\mathcal{I}| \le C$ and

$$|\varpi_j - \xi_*' G^{\mathcal{I},\mathcal{I}}\xi_*| \equiv |\xi_*' Q_1 \xi_* - \xi_*' G^{\mathcal{I},\mathcal{I}}\xi_*| \le C\|G^{\mathcal{I},\mathcal{I}} - Q_1\| = o(1). \tag{5.157}$$

Combining (5.156) and (5.157) gives (5.155). □

## APPENDIX A: SUPPLEMENTARY PROOFS

In this section, we prove Lemma 5.5, 5.7, 5.8 and 5.11.

**A.1. Proof of Lemma 5.5.** Write $\overline{\kappa}_p = \max_{\log(p)\le j\le p-\log(p)} G^{-1}(j,j)$ and $a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f^{-1}(\omega)d\omega$. The assertion of Lemma 5.5 is

$$\lim_{p\to\infty} \overline{\kappa}_p = a_0.$$

To show this, denote $\underline{\kappa}_p = \min_{\log(p)\le j\le p-\log(p)} G^{-1}(j,j)$, and $\kappa_p = \mathrm{trace}(G^{-1})/p$. Since $\log(p) \ll p$ and all diagonals of $G^{-1}$ are bounded from above, it follows from definitions that

$$\underline{\kappa}_p + o(1) \le \kappa_p \le \overline{\kappa}_p + o(1). \tag{A.158}$$

At the same time, the conditions of Lemma 1.5 ensure that $f^*(\omega)$ is continuously differentiable on $[-\pi,\pi]$. By [26],

$$\lim_{p\to\infty} \kappa_p = a_0.$$

Therefore, $\mathrm{liminf}_{p\to\infty}\overline{\kappa}_p \ge \lim_{p\to\infty}\kappa_p = a_0$, and all we need to show is $\mathrm{limsup}_{p\to\infty}\overline{\kappa}_p \le a_0$.

Towards this end, write $G = G_p$ to emphasize on its dependence of $p$. For any positive definite $p\times p$ matrix $A$ and a subset $V \subset \{1,\cdots,p\}$, if we let $B_1$ be the inverse of $A^{V,V}$ and $B_2$ the $(V,V)$-block of $A^{-1}$, then by elementary algebra, $B_2 - B_1$ is positive semi-definite. Now, for any $(i,j)$ such that $\log(p) < j < p - \log(p)$ and $1 \le i \le \lfloor\log(p)\rfloor$, let $V = \{j - i + 1, \cdots, j - i + \lfloor\log(p)\rfloor\}$ ($\lfloor x\rfloor$ denotes the largest integer $k$ such that $k \le x$). Applying the above argument to the set $V$ and matrix $A = G_p$, we have $[(G_p)^{V,V}]^{-1}(i,i) \le G_p^{-1}(j,j)$. At the same time, the Toeplitz structure yields $(G_p)^{V,V} = G_{\lfloor\log(p)\rfloor}$. As a result, $G^{-1}_{\lfloor\log(p)\rfloor}(i,i) \le G_p^{-1}(j,j)$. Since this holds for all $i$ and $j$, we have

$$\overline{\kappa}_{\lfloor\log(p)\rfloor} \le \underline{\kappa}_p.$$

Combining this with the first inequality of (A.158), $\overline{\kappa}_{\lfloor\log(p)\rfloor} \le \kappa_p + o(1)$. It follows that $\mathrm{limsup}_{p\to\infty}\overline{\kappa}_p \le \lim_{p\to\infty}\kappa_p$ and the claim follows.

We remark that additionally $\lim_{p\to\infty}\underline{\kappa}_p = a_0$, whose proof is similar so we omit. □

**A.2. Proof of Lemma 5.7.** Fix $\log(p) \leq j \leq p - \log(p)$. Denote the collection of pairs of sets

$$\mathcal{C}_j = \big\{(F, N) : \min(F \cup N) = j,\ F \cap N = \emptyset,\ F \neq \emptyset\big\},$$

and its sub-collection

$$\mathcal{C}_j^* = \big\{(F, N) \in \mathcal{C}_j : F \in \mathcal{R}_p,\ (F \cup N) \in \mathcal{R}_p,\ |F| \leq 3 \text{ and } |N| \leq 2\big\},$$

where we recall that $\mathcal{R}_p$ is the collection of sets that are formed by consecutive nodes. The claim now reduces to

$$\min_{(F,N)\in\mathcal{C}_j^*} \psi(F, N) = \min_{(F,N)\in\mathcal{C}_j} \psi(F, N) + o(1).$$

Noting that $\mathcal{C}_j^* \subset \mathcal{C}_j$, it suffices to show for any $(F, N) \in \mathcal{C}_j$, there exists $(F', N')$ such that

$$\psi(F', N') \leq \psi(F, N) + o(1) \quad \text{and} \quad (F', N') \in \mathcal{C}_j^*. \tag{A.159}$$

To show (A.159), we introduce the notation $(F', N') \preceq (F, N)$ to indicate

$$\psi(F', N') \leq \psi(F, N), \quad |F'| \leq |F|, \quad \text{and} \quad |N'| \leq |N|.$$

Using these notations, we claim:

(a) For any $(F, N) \in \mathcal{C}_j$, there exists $(F', N') \in \mathcal{C}_j$ such that $\psi(F', N') \leq \psi(F, N) + o(1)$ and $|F'| \leq 3$.
(b) For any $(F, N) \in \mathcal{C}_j$, there exists $(F', N') \in \mathcal{C}_j$ such that $(F', N') \preceq (F, N)$ and $(F' \cup N') \in \mathcal{R}_p$.
(c) For any $(F, N) \in \mathcal{C}_j$ satisfying $(F \cup N) \in \mathcal{R}_p$, there exists $(F', N') \in \mathcal{C}_j$ such that $(F', N') \preceq (F, N)$, $(F' \cup N') \in \mathcal{R}_p$ and $F' \in \mathcal{R}_p$.
(d) For any $(F, N) \in \mathcal{C}_j$ satisfying $(F \cup N) \in \mathcal{R}_p$ and $F \in \mathcal{R}_p$, there exists $(F', N') \in \mathcal{C}_j$ such that $(F', N') \preceq (F, N)$, $(F' \cup N') \in \mathcal{R}_p$, $F' \in \mathcal{R}_p$ and $|N'| \leq 2$.

Now, for any $(F, N) \in \mathcal{C}_j$, we construct $(F', N')$ as follows: First, by (a), there exists $(F_1, N_1)$ such that $\psi(F_1, N_1) \leq \psi(F, N) + o(1)$, and $|F_1| \leq 3$. Second, by (b) and (c), there exists $(F_2, N_2)$ such that $(F_2, N_2) \preceq (F_1, N_1)$, $F_2 \in \mathcal{R}_p$ and $(F_2 \cup N_2) \in \mathcal{R}_p$. Finally, by (d), there exists $(F_3, N_3)$ such that $(F_3, N_3) \preceq (F_2, N_2)$, $(F_3 \cup N_3) \in \mathcal{R}_p$, $F_3 \in \mathcal{R}_p$ and $|N_3| \leq 2$. Let $(F', N') = (F_3, N_3)$.

By the construction, $(F' \cup N') \in \mathcal{R}_p$, $F' \in \mathcal{R}_p$ and

$$\psi(F', N') = \psi(F_3, N_3) \leq \psi(F_2, N_2) \leq \psi(F_1, N_1) \leq \psi(F, N) + o(1).$$

Moreover, $|F'| = |F_3| \leq |F_2| \leq |F_1| \leq 3$, and $|N'| = |N_3| \leq 2$. So $(F', N')$ satisfies (A.159).

All remains is to verify the claims (a)-(d). We need the following results, which follow from basic algebra and we omit the proof: First, recall the definition of $\omega(F,N)$ in (1.34). For any fixed $(F,N)$, let $\mathcal{I} = F \cup N$ and $R = (G^{\mathcal{I},\mathcal{I}})^{-1}$. Then

$$\omega(F,N) = \min_{\xi \in \mathbb{R}^{|F|}: |\xi_i| \geq 1} \xi'(R^{F,F})^{-1}\xi. \tag{A.160}$$

Second, when $(F \cup N) \in \mathcal{R}_p$,

$$R = \frac{1}{j}\eta\eta' + \Sigma_*^{(k)}, \qquad k = |F \cup N|, \tag{A.161}$$

where $\eta = (1, 0, \cdots, 0)'$ and $\Sigma_*^{(k)}$ is as in (5.110).

Now, we show (a). The case $|F| \leq 3$ is trivial, so without loss of generality we assume $|F| \geq 4$. Take

$$F' = \{j+1, j+2\}, \qquad N' = \{j\}.$$

We check that $(F', N')$ satisfies the requirement in (a). It is obvious that $(F', N') \in \mathcal{C}_j$ and $|F'| \leq 3$. We only need to check $\psi(F', N') \leq \psi(F, N) + o(1)$. On one hand, direct calculations yield $\omega(F', N') = (j+1)/(j+2) = 1 + o(1)$, and

$$\psi(F', N') \leq 2\vartheta + r/4 + o(1).$$

On the other hand, by (A.160), $\omega(F,N) \geq |F| \cdot [\lambda_{\max}(R)]^{-1} \geq |F| \cdot \lambda_{\min}(G)$. Noting that $G^{-1} = H$, we have $\|G^{-1}\| \leq \|H\|_\infty \leq 4$. So $\lambda_{\min}(G) \geq 1/4$. Therefore, $\omega(F,N) \geq 1$. It follows that

$$\psi(F,N) \geq |F|\vartheta/2 + \omega(F,N)r/4 \geq 2\vartheta + r/4.$$

Combining the two parts, we have $\psi(F', N') \leq \psi(F, N) + o(1)$.

Next, we verify (b). We construct $(F', N')$ by constructing a sequence of $(F^{(t)}, N^{(t)})$ recursively: Initially, set $F^{(1)} = F$ and $N^{(1)} = N$. On round $t$, write $F^{(t)} \cup N^{(t)} = \{j_1, \cdots, j_k\}$, where the nodes are arranged in the acceding order and $k = |F^{(t)} \cup N^{(t)}|$. Let $l_0$ be the largest index such that $j_l = j_1 + l - 1$ for all $l \leq l_0$. If $l_0 = k$, then the process terminates. Otherwise, let $L = j_{l_0+1} - j_1 - l_0$ and update

$$F^{(t+1)} = \big\{j_l - L \cdot 1\{l > l_0\} : j_l \in F^{(t)}\big\}, \quad N^{(t+1)} = \big\{j_l - L \cdot 1\{l > l_0\} : j_l \in N^{(t)}\big\}.$$

By the construction, it is not hard to see that $l_0$ strictly increases as $t$ increases, and $k$ remains unchanged. So the process terminates in finite

rounds. Let $T$ be the number of rounds when the process terminates, we construct $(F', N')$ by

$$F' = F^{(T)}, \qquad N' = N^{(T)}.$$

Now, we justify that $(F', N')$ satisfies the requirement in (b). First, it is seen that $\min(F^{(t)} \cup N^{(t)}) = j$ on every round $t$. So $\min(F \cup N) = j$ and $(F, N) \in \mathcal{C}_j$. Second, on round $T$, $l_0 = k$, which implies $(F' \cup N') \in \mathcal{R}_p$. Third, $|F^{(t)}|$ and $|N^{(t)}|$ keep unchanged as $t$ increases, so $|F'| = |F|$ and $|N'| = |N|$. Finally, it remains to check $\psi(F', N') \leq \psi(F, N)$. It suffices to show

$$\psi(F^{(t+1)}, N^{(t+1)}) \leq \psi(F^{(t)}, N^{(t)}), \quad \text{for } t = 1, \cdots, T-1. \tag{A.162}$$

Let $\mathcal{I} = F^{(t)} \cup N^{(t)}$ and $\mathcal{I}_1 = F^{(t+1)} \cup N^{(t+1)}$. We observe that $G^{\mathcal{I}_1, \mathcal{I}_1} = G^{\mathcal{I}, \mathcal{I}} - L\eta\eta'$, where $\eta = (0'_{l_0}, 1'_{k-l_0})'$. So $G^{\mathcal{I},\mathcal{I}} - G^{\mathcal{I}_1, \mathcal{I}_1}$ is positive semi-definite. It follows form (A.160) that $\omega(F^{(t+1)}, N^{(t+1)}) \leq \omega(F^{(t)}, N^{(t)})$, and hence (A.162) holds by recalling that $|F^{(t+1)}| = |F^{(t)}|$ and $|N^{(t+1)}| = |N^{(t)}|$.

Third, we prove (c). By assumptions, $(F \cup N) \in \mathcal{R}_p$, so that we can write $F \cup N = \{j, j+1, \cdots, j+k\}$, where $k + 1 = |F \cup N|$. The case $F \in \mathcal{R}_p$ is trivial. In the case $F \notin \mathcal{R}_p$, we construct $(F', N')$ as follows: Let $i_0$ be the smallest index such that $i_0 \notin F$ and both $F_1 = F \cap \{i : i < i_0\}$ and $F_2 = F \backslash F_1$ are not empty. We note that such $i_0$ exists because $F \notin \mathcal{R}_p$. Let

$$F' = F_1 = \{i \in F : i < i_0\}, \qquad N' = \{i \in N : i \leq i_0\}.$$

To check that $(F', N')$ satisfies the requirement in (c), first note that $\min(F' \cup N') = j$ and hence $(F', N') \in \mathcal{C}_j$. Second, it is easy to see that $|F'| \leq |F|$ and $|N'| \leq |N|$. Third, from the definition of $i_0$, $F' \in \mathcal{R}_p$. Additionally, since $i_0 \in N$, we have $F' \cup N' = \{j, j+1, \cdots, i_0\} \in \mathcal{R}_p$. Last, we check $\psi(F', N') \leq \psi(F, N)$: Since $|F'| \leq |F|$ and $|N'| \leq |N|$, it suffices to show

$$\omega(F', N') \leq \omega(F, N). \tag{A.163}$$

Write $\mathcal{I} = F \cup N$ and denote $R = (G^{\mathcal{I},\mathcal{I}})^{-1}$. From (A.161), $R$ is tri-diagonal. So $R^{F,F}$ is block-diagonal in the partition $F = F_1 \cup F_2$. Using (A.160), it is easy to see

$$\omega(F_1, \mathcal{I} \backslash F_1) \leq \omega(F, \mathcal{I} \backslash F) \equiv \omega(F, N).$$

At the same time, notice that both $\mathcal{I}$ and $\mathcal{I}' = F' \cup N'$ have the form $\{j, j+1, \cdots, m\}$ with $m \geq \max(F_1) + 1$. Applying (A.160) and (A.161), by direct calculations,

$$\omega(F_1, \mathcal{I} \backslash F_1) = \omega(F_1, \mathcal{I}' \backslash F_1) \equiv \omega(F', N').$$

Combining the two parts gives (A.163).

Finally, we justify (d). By assumptions, $(F \cup N) \in \mathcal{R}_p$ and $F \in \mathcal{R}_p$, so that we write $F \cup N = \{j, j+1, \cdots, k\}$, and $F = \{j_0, j_0+1, \cdots, k_0\}$, where $j_0 \geq j$ and $k_0 \leq k$. The case $|N| \leq 2$ is trivial. In the case $|N| > 2$, let $m_0 = |F|$ and we construct $(F', N')$ as follows:

$$\begin{array}{lr} F' = F, \quad N' = \{k_0+1\}, & \text{when } j_0 = j; \\ F' = \{j+1, j+2, \cdots, j+m_0\}, \quad N' = \{j,\ j+m_0+1\}, & \text{when } j_0 > j, k_0 < k; \\ F' = \{j+1, j+2, \cdots, j+m_0\}, \quad N' = \{j\}, & \text{when } j_0 > j, k_0 = k. \end{array}$$

Now, we show that $(F', N')$ satisfies the requirement in (d). First, by the construction, $(F', N') \in \mathcal{C}_j$, $(F' \cup N') \in \mathcal{R}_p$ and $F' \in \mathcal{R}_p$. Second, $|F'| = |F|$, $|N'| \leq 2 < |N|$. Third, we check $\psi(F', N') \leq \psi(F, N)$. Applying (A.160) and (A.161), direct calculations yield $\omega(F', N') = \omega(F, N)$. This, together with $|F'| \leq |F|$ and $|N'| \leq |N|$, proves $\psi(F', N') \leq \psi(F, N)$. □

**A.3. Proof of Lemma 5.8.** Recalling the definition of $\mathcal{C}_j^*$ in the proof of Lemma 5.7, the claim reduces to

$$\min_{(F,N)\in\mathcal{C}_j^*} \psi(F, N) = \min_{(F,N)\in\mathcal{C}_1^*} \psi^{(\infty)}(F, N) + o(1), \qquad \log(p) \leq j \leq p - \log(p).$$

We argue that on both sides, the minimum is not attained on $(F, N)$ such that $|N| = 0$ and $|F| = 1$. In this case, on the left hand side, $F = \{j\}$ and $N = \emptyset$. By direct calculations, $\omega(F, N) = j \geq \log(p)$, and hence $\psi(F, N)$ can not be the minimum. Similarly, on the right hand side, $\omega^{(\infty)}(F, N) = \infty$ by definition, and the same conclusion follows. Therefore, the claim is equivalent to

$$\min_{(F,N)\in\mathcal{C}_j^*:|F|+|N|>1} \psi(F, N) = \min_{(F,N)\in\mathcal{C}_1^*:|F|+|N|>1} \psi^{(\infty)}(F, N) + o(1).$$

Fix $\log(p) \leq j \leq p - \log(p)$. Define a one-to-one mapping from $\mathcal{C}_j^*$ to $\mathcal{C}_1^*$, where given any $(F, N) \in \mathcal{C}_j^*$, it is mapped to $(F_1, N_1)$ such that

$$F_1 = \{i - j + 1 : i \in F\}, \qquad N_1 = \{i - j + 1 : i \in N\}.$$

To show the claim, it suffices to show when $|F| + |N| > 1$,

$$\psi(F, N) = \psi^{(\infty)}(F_1, N_1) + o(1).$$

Since $|F_1| = |F|$ and $|N_1| = |N|$, it is sufficient to show

$$\omega(F, N) = \omega^{(\infty)}(F_1, N_1) + o(1). \tag{A.164}$$

Now, we show (A.164). Consider the case $N \neq \emptyset$ first. Suppose $|\mathcal{I}| = k$ and write $\mathcal{I} = F \cup N = \{j, \cdots, j+k-1\}$, where $1 < k \leq 5$. Let $R = (G^{\mathcal{I},\mathcal{I}})^{-1}$ and $R_* = (\Sigma_*^{(k)})^{F_1,F_1}$, where $\Sigma_*^{(k)}$ is defined in (5.110). We note that when $N \neq \emptyset$, $R_*$ is invertible. Using (A.160) and the definition of $\omega^{(\infty)}$,

$$|\omega(F,N) - \omega^{(\infty)}(F_1,N_1)| \leq \max_{\xi \in \mathbb{R}^k : |\xi_i| \leq 2a} \left|\xi'[(R^{F,F})^{-1} - R_*^{-1}]\xi\right|. \tag{A.165}$$

Since $\mathcal{I} \in \mathcal{R}_p$, we apply (A.161) and obtain

$$R^{F,F} = \frac{1}{j}(\eta^{F_1})(\eta^{F_1})' + (\Sigma_*^{(k)})^{F_1,F_1},$$

where $\eta = (1, 0, \cdots, 0)' \in \mathbb{R}^k$. By matrix inverse formula,

$$\xi'[(R^{F,F})^{-1} - R_*^{-1}]\xi = -[j + (\eta^{F_1})' R_*^{-1} \eta^{F_1}]^{-1} (\xi' R_*^{-1} \eta^{F_1})^2. \tag{A.166}$$

Combining (A.165) and (A.166),

$$|\omega(F,N) - \omega^{(\infty)}(F_1,N_1)| \leq j^{-1} \max_{\xi \in \mathbb{R}^k : |\xi_i| \leq 2a} |\xi' R_*^{-1} \eta^{F_1}|^2 \leq j^{-1} \cdot C \|R_*^{-1}\|^2.$$

Since $N_1 \neq \emptyset$ and $k$ is finite, $\lambda_{\min}(R_*) \geq C > 0$ and hence $\|R_*^{-1}\| \leq C$. Noting that $j \geq \log(p)$, (A.164) follows directly.

Next, consider the case $N = \emptyset$. Suppose $|F| = k$ and write $F = \{j, \cdots, j + k - 1\}$, where $1 < k \leq 3$. We observe that $G^{F,F} = j11' + \Omega_*^{(k)}$, where $\Omega_*^{(k)}$ is defined in (5.111). By definition

$$\omega(F,N) = \min_{\xi \in \mathbb{R}^k : |\xi_i| \geq 1} \xi' G^{F,F} \xi = \min_{\xi \in \mathbb{R}^k : |\xi_i| \geq 1} \left[ j(1'\xi)^2 + \xi' \Omega_*^{(k)} \xi \right].$$

On one hand, if we let $\xi^*$ be one minimizer in the definition of $\omega^{(\infty)}(F_1,N_1)$, then $1'\xi^* = 0$. As a result,

$$\omega(F,N) \leq j(1'\xi^*)^2 + (\xi^*)' \Omega_*^{(k)} \xi^* = (\xi^*)' \Omega_*^{(k)} \xi^* \equiv \omega^{(\infty)}(F_1,N_1). \tag{A.167}$$

On the other hand, we can show

$$\omega(F,N) \geq \omega^{(\infty)}(F_1,N_1) - 1/(j+1). \tag{A.168}$$

Combing (A.167) and (A.168), and noting that $j \geq \log(p)$, we obtain (A.164).

It remains to show (A.168). When $k = 2$, by direct calculations, $\omega(F,N) = \omega^{(\infty)}(F,N) = 1$. When $k > 2$, write $\xi = (\xi_1, \xi_2, \tilde{\xi}')'$ for any $\xi \in \mathbb{R}^k$, and

introduce the function $g(x) = \sum_{i=1}^{k-2}(x_i + x_{i+1} + \cdots + x_{k-2})^2$, for $x \in \mathbb{R}^{k-2}$. We observe that

$$\xi'\Omega_*^{(k)}\xi = (1'\xi - \xi_1)^2 + g(\tilde{\xi}). \tag{A.169}$$

Let $g_{\min} = \min_{x\in\mathbb{R}^{k-2}:|x_i|\geq 1} g(x)$. We claim that there exists $q \in \mathbb{R}^k$ such that

$$1'q = 0, \qquad q'\Omega_*^{(k)}q = 1 + g_{\min}, \qquad \text{and} \qquad |q_i| \geq 1, \quad \text{for } 1 \leq i \leq k.$$

To see this, note that under the constraints $|x_i| \geq 1$, $g(x)$ is obviously minimized at $x^* = (\cdots, -1, 1, -1, 1)$. Observing that $1'(x^*)$ is either 0 or 1, we let $q = (1, -1, (x^*)')'$ when $1'(x^*) = 0$, and let $q = (1, -2, (x^*)')'$ when $1'(x^*) = 1$. Using (A.169), it is easy to check that $q$ satisfies the above requirements. It follows that

$$\omega^{(\infty)}(F_1, N_1) = \min_{\xi\in\mathbb{R}^k:|\xi_i|\geq 1, 1'\xi=0} \xi'\Omega_*^{(k)}\xi \leq q'\Omega_*^{(k)}q = 1 + g_{\min}. \tag{A.170}$$

At the same time, since $G^{F,F} = j11' + \Omega_*^{(k)}$, we can write from (A.169) that

$$\xi'G^{F,F}\xi = j(1'\xi)^2 + (1'\xi - \xi_1)^2 + g(\tilde{\xi}). \tag{A.171}$$

Note that $\min_y\{jy^2 + (y - c)^2\} = c^2\ j/(j + 1)$, for any $c \in \mathbb{R}$. So

$$j(1'\xi)^2 + (1'\xi - \xi_1)^2 \geq |\xi_1|^2 j/(j + 1).$$

Plugging this into (A.171), we find that

$$\omega(F, N) = \min_{\xi\in\mathbb{R}^k:|\xi_i|\geq 1} \xi'G^{F,F}\xi \geq j/(j + 1) + g_{\min}. \tag{A.172}$$

Combining (A.170) and (A.172) gives (A.168). □

**A.4. Proof of Lemma 5.11.** To show the claim, we first introduce a key lemma: Fix a linear filter $D_{h,\eta}$, for any dimension $k > h$, let $\tilde{D}^{(k)}$ be the $(k - h) \times k$ matrix, where for each $1 \leq i \leq k - h$, $\tilde{D}^{(k)}(i, i) = 1$, $\tilde{D}^{(k)}(i, i + 1) = \eta_1$, $\cdots$, $\tilde{D}^{(k)}(i, i + h) = \eta_h$, and $\tilde{D}^{(k)}(i, j) = 0$ for other $j$. Define the null space of $D_{h,\eta}$ in dimension $k$, $Null_k(\eta)$, as the collection of all vectors $\xi \in \mathbb{R}^k$ that satisfies $\tilde{D}^{(k)}\xi = 0$. The following lemma is proved below.

LEMMA A.1. *For a given $\eta$, if RCA holds, then for sufficiently large $n$ and any $k \geq n$, there exists an orthonormal basis of $Null_k(\eta)$, denoted as $\xi^{(1)}, \cdots, \xi^{(h)}$, such that*

$$\max_{1\leq i\leq k-n, 1\leq j\leq h} |\xi_i^{(j)}|^2 \leq C_\eta n^{-1},$$

*where $C_\eta > 0$ is a constant that only depends on $\eta$.*

Second, we state some observations. Fix $\mathcal{I} \trianglelefteq \mathcal{G}^+$. Partition $\mathcal{I}^{pe}$ uniquely as $\mathcal{I}^{pe} = \cup_{t=1}^T V_t$, so that $V_t = \{i_t, i_t + 1, \cdots, j_t - 1, j_t\}$ is formed by consecutive nodes and $j_t < i_{t+1}$ for all $t$. Denote $M = |\mathcal{J}^{pe}| - |\mathcal{I}^{pe}|$. It is easy to see that $T \leq M$ and $M \leq h|\mathcal{I}| \leq l_0 h$, so both $M$ and $T$ are finite. Let $\tilde{V}_t = \{1 \leq j \leq p : D(i,j) \neq 0 \text{ for some } i \in V_t\}$ and define $Null(V_t, \tilde{V}_t)$ in the same way as $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$. Recall that $\mathcal{F}$ is the mapping from nodes in $\mathcal{J}^{pe}$ to their orders in $\mathcal{J}^{pe}$. Similarly, define the mapping $\mathcal{F}_t$ from $\tilde{V}_t$ to $\{1, \cdots, |\tilde{V}_t|\}$ that maps each $j \in \tilde{V}_t$ to its order in $\tilde{V}_t$. Denote $\mathcal{I}_t = \mathcal{F}_t(\mathcal{I} \cap V_t)$. We observe that:

- $(O_1)$ $\tilde{V}_t \cap \tilde{V}_{t'} \neq \emptyset$ only when $|t - t'| \leq 1$; and $|\tilde{V}_t \cap \tilde{V}_{t+1}| \leq h - 1$, for all $t$.
- $(O_2)$ $Null(V_t, \tilde{V}_t) = Null_{|\tilde{V}_t|}(\eta)$ for all $t$, where $Null_k(\eta)$ is as in Lemma A.1.
- $(O_3)$ $\mathcal{J}^{pe} = \cup_{t=1}^T \tilde{V}_t$; and $|\tilde{V}_t| \geq |V_t| \geq 2\ell^{pe} + 1$, for all $t$.
- $(O_4)$ Any node $i \in \mathcal{I}_t$ satisfies that $1 \leq i < |\tilde{V}_t| - \ell^{pe}$, for all $t$.
- $(O_5)$ For any $\xi \in \mathbb{R}^{|\mathcal{J}^{pe}|}$, $\xi \in Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$ if and only if $\xi^{\mathcal{F}(\tilde{V}_t)} \in Null(V_t, \tilde{V}_t)$ for all $t$, where $\xi^{\mathcal{F}(\tilde{V}_t)}$ is the subvector of $\xi$ formed by elements in $\mathcal{F}(\tilde{V}_t)$.

Due to $(O_2)$ and Lemma A.1, for each $t$, there exists an orthonormal basis $\xi^{(t,1)}, \cdots, \xi^{(t,h)}$ for $Null(V_t, \tilde{V}_t)$ such that

$$\max_{1\leq i\leq |\tilde{V}_t|-n, 1\leq j\leq h} |\xi_i^{(t,j)}|^2 \leq C_\eta n^{-1}, \quad \text{for any } 1 \leq n < |\tilde{V}_t|. \tag{A.173}$$

Let $U_t$ be the matrix formed by the last $h$ rows of $[\xi^{(t,1)}, \cdots, \xi^{(t,h)}]$. From the explicit form of the basis in the proof of Lemma A.1, we further observe:

- $(O_6)$ $c' \leq \lambda_{\min}(U_t U_t') \leq \lambda_{\max}(U_t U_t') \leq 1 - c$, where $0 < c, c' < 1$ and $c + c' < 1$.
- $(O_7)$ For each $1 \leq h_0 \leq h$, the submatrix of $U_t$ formed by its last $h_0$ rows has a rank $h_0$.

Now, we show the claim by constructing a matrix $W$, whose columns form an orthonormal baisis for $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$, and it satisfies

$$\max_{1\leq i\leq |\mathcal{J}^{pe}|-n, 1\leq j\leq M} |W(i,j)|^2 \leq C n^{-1}, \quad \text{for any } 1 \leq n < |\mathcal{J}^{pe}|. \tag{A.174}$$

In fact, once such $W$ is constructed, any $U$ whose columns form an orthonromal baisis for $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$ can be written as

$$U = WR,$$

where $R$ has the dimension $M \times M$ and $R'R$ is an identity matrix. By basic algebra, for any $m \times n$ matrix $A$ and $n \times p$ matrix $B$, $\max_{1 \leq j \leq p} |(AB)(i,j)|^2 \leq n\|B'B\| \cdot \max_{1 \leq k \leq n} |A(i,k)|^2$ for each $1 \leq i \leq m$. Applying this to $W$ and $R$, and noting that $\|R'R\| = 1$ and that $M$ is finite, we obtain

$$\max_{i \in \mathcal{F}(\mathcal{I}), 1 \leq j \leq M} |U(i,j)|^2 \leq C \max_{i \in \mathcal{F}(\mathcal{I}), 1 \leq j \leq M} |W(i,j)|^2. \tag{A.175}$$

At the same time, for any $i \in \mathcal{I}$, there exists a unique $t$ such that $i \in \mathcal{I} \cap V_t$. In addition, from $(O_4)$, $\mathcal{F}_t(i) < |\tilde{V}_t| - \ell^{pe}$. By the construction, this implies $\mathcal{F}(i) < |\mathcal{J}^{pe}| - \ell^{pe}$. Combining this to (A.174), we find that

$$\max_{i \in \mathcal{F}(\mathcal{I}), 1 \leq j \leq M} |W(i,j)|^2 \leq C(\ell^{pe})^{-1}. \tag{A.176}$$

The claim then follows from (A.175) and (A.176).

To construct $W$, the key is to recursively construct matrices $W_T, W_{T-1}, \cdots, W_1$. Denote $m_t = h - |\tilde{V}_t \cap \tilde{V}_{t+1}|$, with $m_T = h$ by convention; $M_t = \sum_{s=t}^T m_s$ and $L_t = |\cup_{s=t}^T \tilde{V}_s|$; in particular, $M_1 = |\mathcal{J}^{pe}| - |\mathcal{I}^{pe}| = M$ and $L_1 = |\mathcal{J}^{pe}|$. Initially, construct the $L_T \times M_T$ matrix

$$W_T = \left[\xi^{(T,1)}, \cdots, \xi^{(T,h)}\right],$$

where $\{\xi^{(T,j)} : 1 \leq j \leq h\}$ is the orthonormal basis in (A.173). Given $W_{t+1}$, construct the $L_t \times M_t$ matrix $W_t$ as follows: Denote $\widetilde{W}_{t+1}$ the submatrix of $W_{t+1}$ formed by its first $|\tilde{V}_t \cap \tilde{V}_{t+1}|$ $(= h - m_t)$ rows and write

$$\left[\xi^{(t,1)}, \cdots, \xi^{(t,h)}\right] = \begin{bmatrix} A_t \\ B_t \end{bmatrix},$$

where $A_t$ has $(|\tilde{V}_t| - h - m_t)$ rows and $B_t$ has $(h - m_t)$ rows. From $(O_7)$, the rank of $B_t$ is $(h - m_t)$. Hence, there exists an $h \times m_t$ matrix $Q_t$, such that $Q_t'Q_t$ is an identity matrix and $B_t Q_t = 0$. Now, construct

$$W_t = \begin{bmatrix} A_t B_t'(B_t B_t')^{-1} \widetilde{W}_{t+1} & A_t Q_t \\ W_{t+1} & 0 \end{bmatrix}. \tag{A.177}$$

Continue this process until we obtain $W_1$ and let

$$W = W_1 (W_1' W_1)^{-1/2}.$$

Below, we check that $W$ satisfies the requirement. First, we show that the columns of $W$ form an orthonormal basis of $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$. Since $W$ has $M = |\mathcal{J}^{pe}| - |\mathcal{I}^{pe}|$ columns and its columns are orthonormal, it suffices to show that all its columns belong to $Null(\mathcal{I}^{pe}, \mathcal{J}^{pe})$. By $(O_5)$, we only need to show that for each $1 \leq t \leq T$, in the submatix of $W$ formed by restricting rows into $\mathcal{F}(\tilde{V}_t)$, all its columns belong to $Null(V_t, \tilde{V}_t)$. By the construction, only the first $M_t$ columns of this submatrix are non-zero and they are equal to

$$\begin{bmatrix} A_t B_t'(B_t B_t')^{-1}\widetilde{W}_{t+1} & A_t Q_t \\ \widetilde{W}_{t+1} & 0 \end{bmatrix} = \begin{bmatrix} A_t \\ B_t \end{bmatrix} \begin{bmatrix} B_t'(B_t B_t')^{-1}\widetilde{W}_t, \ Q_t \end{bmatrix},$$

where in the equality we have used the facts that $\widetilde{W}_t = B_t B_t'(B_t B_t')^{-1}\widetilde{W}_t$ and $B_t Q_t = 0$. Combining this to the definition of $A_t$ and $B_t$, we find that each column of the above matrix is a linear combination of $\{\xi^{(t,1)}, \cdots, \xi^{(t,h)}\}$ and hence belongs to $Null(V_t, \tilde{V}_t)$.

Second, we show that $W$ satisfies (A.174). It suffices to show, for $t = T, \cdots, 1$,

(a) $\max_{1 \leq i \leq L_t - n, 1 \leq j \leq M_t} |W_t(i,j)|^2 \leq Cn^{-1}$, for any $1 \leq n < L_t$.
(b) $\lambda_{\min}(W_t' W_t) \geq C > 0$.

In fact, once (a) and (b) are proved, by taking $t = 1$ and noticing that $L_1 = |\mathcal{J}^{pe}|$, we have $\max_{1 \leq i \leq |\mathcal{J}^{pe}| - n, 1 \leq j \leq M} |W_1(i,j)|^2 \leq Cn^{-1}$, for $1 \leq n < |\mathcal{J}^{pe}|$; and $\|(W_1'W_1)^{-1}\| = [\lambda_{\min}(W_1'W_1)]^{-1} \leq C$. Hence, by similar arguments in (A.175), for each $1 \leq i \leq |\mathcal{J}^{pe}| - n$, $\max_{1 \leq j \leq M} |W(i,j)|^2 \leq M\|(W_1'W_1)^{-1}\| \cdot \max_{1 \leq j \leq M} |W_1(i,j)|^2 \leq Cn^{-1}$. This gives (A.174).

It remains to show (a) and (b). Note that for $W_T$, by the construction and (A.173), (a) and (b) hold trivially. We aim to show that if (a) and (b) hold for $W_{t+1}$, then they also hold for $W_t$. For preparation, we argue that

$$\|A_t B_t'(B_t B_t')^{-1}\widetilde{W}_{t+1}\|^2 \leq C(\ell^{pe})^{-1} = o(1). \tag{A.178}$$

To see this, note that $L_{t+1} \geq 2\ell^{pe} + 1$ from $(O_3)$; in particular, $h - m_t \ll L_{t+1} - \ell^{pe}$. Hence, if (a) holds for $W_{t+1}$, $\max_{1 \leq i \leq h - m_t, 1 \leq j \leq M_{t+1}} |W_{t+1}(i,j)|^2 \leq C(\ell^{pe})^{-1}$, i.e., $|\widetilde{W}_{t+1}(i,j)| \leq C(\ell^{pe})^{-1}$, for any $(i,j)$. Since $\widetilde{W}_{t+1}$ has a finite dimension, this yields $\|\widetilde{W}_{t+1}\|^2 \leq C(\ell^{pe})^{-1}$. Furthermore, from $(O_6)$ and that $B_t B_t'$ is a submatrix of $U_t U_t'$, $\lambda_{\min}(B_t B_t') \geq c' > 0$. So $\|(B_t B_t')^{-1}\| \leq C$. In addition, $\|A_t\|, \|B_t\| \leq 1$. Combining the above gives (A.178).

Consider (a) first. By (A.177), (A.178) and the assumption on $W_{t+1}$, it suffices to show

$$\max_{1 \leq i \leq |\tilde{V}_t| - n, 1 \leq j \leq m_t} |A_t Q_t(i,j)|^2 \leq Cn^{-1}, \qquad \text{for any } 1 \leq n < |\tilde{V}_t|. \tag{A.179}$$

By similar arguments in (A.175) and the fact that $\|Q_t'Q_t\| = 1$, the left hand side is bounded by $C\max_{1\leq i\leq|\tilde{V}_t|-n, 1\leq j\leq m_t}|A_t(i,j)|^2$. Therefore, (A.179) follows from (A.173) and the definition of $A_t$.

Next, consider (b). Using (A.177) and (A.178), we can write

$$W_t'W_t = \begin{bmatrix} W_{t+1}'W_{t+1} + \Delta_1 & \Delta_2 \\ \Delta_2' & Q_t'A_t'A_tQ_t \end{bmatrix},$$

where $\|\Delta_1\| = o(1)$ and $\|\Delta_2\| = o(1)$. So it suffices to show $\lambda_{\min}(W_{t_1}'W_{t+1}) \geq C$ and $\lambda_{\min}(Q_t'A_t'A_tQ_t) \geq C$. The former follows from the assumption on $W_{t+1}$. To show the latter, note that $Q_t'Q_t$ is an identity matrix, and so $\lambda_{\min}(Q_t'A_t'A_tQ_t) \geq \lambda_{\min}(A_t'A_t)$. Also, since $A_t'A_t + B_t'B_t$ is an identity matrix, $\lambda_{\min}(A_t'A_t) = 1-\lambda_{\max}(B_t'B_t)$. Additionally, $\lambda_{\max}(B_t'B_t) = \lambda_{\max}(B_tB_t')$, where $B_tB_t'$ is a submatrix of $U_tU_t'$, and by $(O_6)$, $\lambda_{\max}(U_tU_t') \leq 1-c$. Combining the above yields $\lambda_{\min}(Q_t'A_t'A_tQ_t) \geq c > 0$. This proves (b). □

A.4.1. *Proof of Lemma A.1.* For each $k \geq h$, we construct a $k\times h$ matrix $U$ whose columns form an orthonormal basis of $Null_k(\eta)$ as follows: Recall the characteristic polynomial $\varphi_\eta(z) = 1 + \eta_1 z + \cdots + \eta_h z^h$. Let $z_1, \cdots, z_m$ be $m$ different roots of $\varphi_\eta(z)$, each replicating $h_1, \cdots, h_m$ times respectively $(h_1 + \cdots + h_m = h)$. For $1 \leq j \leq m$ and $1 \leq s \leq h_j$, when $z_i$ is a real root, let

$$\mu^{(j,s)} = \left(k^{s-1}\frac{1}{z_j^{k-1}},\ \cdots,\ 3^{s-1}\frac{1}{z_j^2},\ 2^{s-1}\frac{1}{z_j},\ 1\right)';$$

and when $z_{j\pm} = |z_j|e^{\pm\sqrt{-1}\theta_j}$, $\theta_j \in (0, \pi/2]$, are a pair of conjugate roots, let

$$\mu^{(j+,s)} = \left(k^{s-1}\frac{\cos(k-1)\theta_j}{|z_j|^{k-1}},\ \cdots,\ 3^{s-1}\frac{\cos 2\theta_j}{|z_j|^2},\ 2^{s-1}\frac{\cos\theta_j}{|z_j|},\ 1\right)',$$
$$\mu^{(j-,s)} = \left(k^{s-1}\frac{\sin(k-1)\theta_j}{|z_j|^{k-1}},\ \cdots,\ 3^{s-1}\frac{\sin 2\theta_j}{|z_j|^2},\ 2^{s-1}\frac{\sin\theta_j}{|z_j|},\ 1\right)'.$$

It is seen that $\{\mu^{(j,s)}, 1 \leq j \leq m, 1 \leq s \leq h_j\}$ are $h$ vectors in $\mathbb{R}^k$. Let $\xi^{(j,s)} = \mu^{(j,s)}/\|\mu^{(j,s)}\|$ for each $(j,s)$, and construct the $k\times h$ matrix

$$R = \left[\xi^{(1,1)}, \cdots, \xi^{(1,h_1)}, \cdots, \xi^{(m,1)}, \cdots, \xi^{(m,h_m)}\right].$$

Define

$$U = R(R'R)^{-1/2}.$$

Now, we show that the vectors $\{\mu^{(j,s)}, 1 \leq j \leq m, 1 \leq s \leq h_j\}$ are linearly independent and span $Null_k(\eta)$. Therefore, $U$ is well defined and its columns

form an orthonormal basis of $Null_k(\eta)$. To see this, note that for any vector $\eta \in \mathbb{R}^k$, if we write $\eta_1 = f(k), \cdots, \eta_k = f(1)$, then $\xi \in Null_k(\eta)$ if and only if $f(i)$'s satisfy the difference equation:

$$f(i) + \eta_1 f(i-1) + \cdots + \eta_h f(i-h) = 0, \quad h+1 \leq i \leq k. \tag{A.180}$$

It is well-known in theories of difference equations that (A.180) has $h$ independent base solutions:

$$f_{j,s}(i) = i^{s-1} z_j^{-i}, \qquad 1 \leq j \leq m, \quad 1 \leq s \leq h_j.$$

By the construction, when $z_j$ is a real root, $\mu^{(j,s)} = (f_{j,s}(k), \cdots, f_{j,s}(1))'$; and when $z_{j\pm}$ are a pair of conjugate roots, $\mu^{(j+,s)}$ and $\mu^{(j-,s)}$ are the real and imaginary parts of the vector $(f_{j,s}(k), \cdots, f_{j,s}(1))'$. So the vectors $\{\mu^{(j,s)}\}$ are linearly independent and they span $Null_k(\eta)$.

Next, we check that the columns of $U$ satisfy the requirement in the claim, i.e., there exists a constant $C_\eta$ such that for any $(n,k)$ satisfying $k \geq n \geq h$,

$$\max_{1 \leq i \leq k-n, 1 \leq j \leq h} |U(i,j)|^2 \leq C_\eta n^{-1}.$$

Since $\max_{1 \leq j \leq h} |U(i,j)| \leq h\|(R'R)^{-1}\| \cdot \max_{1 \leq j \leq h} |R(i,j)|^2$, it suffices to show that

$$\max_{1 \leq i \leq k-n, 1 \leq j \leq h} |R(i,j)|^2 \leq C n^{-1}, \tag{A.181}$$

and that for all $k \geq h$,

$$\lambda_{\min}(R'R) \geq C > 0. \tag{A.182}$$

Consider (A.181) first. It is equivalent to show that

$$\max_{1 \leq i \leq k-n} |\mu_i^{(j,s)}| / \|\mu^{(j,s)}\| \leq C n^{-1/2}, \qquad 1 \leq j \leq m, 1 \leq s \leq h_j. \tag{A.183}$$

In the case $|z_j| > 1$, $\|\mu^{(j,s)}\| \leq C$. In addition, $|z_j|^i \geq C i^{s-1/2}$ for sufficiently large $i$, and hence $\max_{1 \leq i \leq k-n} |\mu_i^{(j,s)}| \leq \max_{i > n} C(i^{s-1} i^{1/2-s}) \leq C n^{-1/2}$. So (A.183) holds. In the case $|z_j| = 1$, it can be shown in analysis that $\|\mu^{(j,s)}\| \geq C k^{s-1/2}$, where $C > 0$ is a constant depending on $\theta_j$ but independent of $k$. Also, $\max_{1 \leq i \leq k-n} |\mu_i^{(j,s)}| \leq \max_{n < i \leq k} C i^{s-1} \leq C k^{s-1}$. Hence, $\max_{1 \leq i \leq k-n} |\mu_i^{(j,s)}| / \|\mu^{(j,s)}\| \leq C k^{-1/2} \leq C n^{-1/2}$ and (A.183) holds.

Next, consider (A.182). $R'R$ is an $h \times h$ matrix. For convenience, we use $\{(j,s) : 1 \le j \le m, 1 \le s \le h_j\}$ to index the entries in $R'R$. By construction, all the diagonals of $R'R$ are equal to 1, and the off-diagonals are equal to

$$(R'R)_{(j,s),(j',s')} = \frac{\langle \mu^{(j,s)}, \mu^{(j',s')} \rangle}{\|\mu^{(j,s)}\| \|\mu^{(j',s')}\|}, \quad (j,s) \neq (j',s'). \tag{A.184}$$

It is easy to see that as $k \to \infty$, each entry of $R'R$ has a finite limit. Therefore, as $k \to \infty$, $R'R$ approaches a fixed $h \times h$ matrix $A$ element-wise. In particular, $\lambda_{\min}(R'R) \to \lambda_{\min}(A)$. Hence, to show (A.182), we only need to prove that $A$ is non-singular.

Write $R = (R_1, R_2)$, where $R_1$ is the submatrix formed by columns corresponding to those roots $|z_j| > 1$, and $R_2$ the submatrix formed by columns corresponding to those roots $|z_j| = 1$. Note that when $|z_j| = 1$ and $|z_{j'}| > 1$, as $k \to \infty$, $|\langle \mu^{(j,s)}, \mu^{(j',s')} \rangle| \le C$, $\|\mu^{(j,s)}\| \to \infty$ and $\|\mu^{(j',s')}\| \ge C$; so $(R'R)_{(j,s),(j',s')} \to 0$. This means $R_1' R_2$ approaches the zero matrix as $k \to \infty$. Consequently,

$$A = \mathrm{diag}(A_1, A_2), \qquad \text{where } R_1'R_1 \to A_1 \text{ and } R_2'R_2 \to A_2, \quad \text{as } k \to \infty.$$

Therefore, it suffices to show that both $A_1$ and $A_2$ are non-singular.

Consider $A_1$ first. Denote $h_0 = \sum_j h_j 1\{|z_j| > 1\}$ so that $R_1$ is a $k \times h_0$ matrix. Let $R_1^*$ be the $k \times h_0$ matrix whose columns are $\{\mu^{(j,s)} : |z_j| > 1\}$, $M$ be the $h_0 \times h_0$ submatrix formed by the last $h_0$ rows of $R^*$ and $\Lambda = \mathrm{diag}(\|\mu^{(j,s)}\|)$ is the $h_0 \times h_0$ diagonal matrix. Now, suppose $A_1$ is singular, i.e., there exists a non-zero vector $b$ such that $b'A_1 b = 0$. This implies $\|R_1 b\| \to 0$ as $k \to \infty$. Using the matrices defined above, we can write $R_1 = R_1^* \Lambda$; so $\|R_1^* \Lambda b\| \to 0$. Since $\|M\Lambda b\| \le \|R_1^* \Lambda b\|$, it further implies $\|M \Lambda b\| \to 0$. First, we observe that $M$ is a fixed matrix independent of $k$. Second, note that when $|z_j| > 1$, $\|\mu^{(j,s)}\| \to c_{js}$, as $k \to \infty$, for some constant $c_{js} > 0$; as a result, $\Lambda \to \Lambda^*$ as $k \to \infty$, where $\Lambda^*$ is a positive definite diagonal matrix. Combining the two parts, $\|M\Lambda b\| \to 0$ implies $\|M(\Lambda^* b)\| = 0$, where $\Lambda^* b$ is a fixed non-zero vector. This means $M$ is singular. Therefore, if we can prove $M$ is non-singular, then by contradiction, $A_1$ is also non-singular.

Now, we show $M$ is non-singular. Let $\widetilde{M}$ be the matrix by re-arranging the rows in $M$ in the inverse order. It is easy to see that $M$ is non-singular if and only if $\widetilde{M}$ is non-singular. For convenience, we use $\{1, \cdots, h_0\} \times \{(j,s) : |z_j| > 1, 1 \le s \le h_j\}$ to index the entries in $\widetilde{M}$. It follows by the construction that

$$\widetilde{M}_{i,(j,s)} = i^{s-1} z_j^{-(i-1)}, \qquad z_j \text{ is a real, } 1 \le i \le h_0$$

$$\left.\begin{array}{l} \widetilde{M}_{i,(j-,s)} = i^{s-1} |z_j|^{-(i-1)} \cos((i-1)\theta_j), \\ \widetilde{M}_{i,(j-,s)} = i^{s-1} |z_j|^{-(i-1)} \sin((i-1)\theta_j), \end{array}\right. \quad z_{j\pm} \text{ are conjugates, } 1 \le i \le h_0.$$

Define an $h_0 \times h_0$ matrix $T$ by

$$T_{i,(j,s)} = i^{s-1} z_j^{-(i-1)}, \quad 1 \le i \le h_0.$$

Let $V$ be the $h_0 \times h_0$ confluent Vandermonde matrices generated by $\{z_j^{-1} : |z_j| > 1\}$:

$$V_{i,(j,s)} = \begin{cases} 0 & 1 \le i \le s-1, \\ \frac{(i-1)!}{(i-s)!} z_j^{-(i-s)} & s \le i \le h_0. \end{cases}$$

First, it is seen that each column of $T$ is a (complex) linear combination of columns in $\widetilde{M}$. Second, we argue that each column of $V$ is a linear combination of columns in $T$. To see this, note that $V_{i,(j,s)}$ can be written in the form $V_{i,(j,s)} = g_{s-1}(i) z_j^{-(i-s)}$, where $g_{s-1}(x) = (x-1)(x-2)\cdots(x-s+1)$ is a polynomial of degree $s-1$. Let $c_0, \cdots, c_{s-1}$ be the coefficients of this polynomial. Then, for each $i > s$, $V_{i,(j,s)} = z_j^{-(i-s)} \sum_{l=0}^{s-1} c_l i^l = \sum_{l=1}^{s} \alpha_l T_{i,(j,l)}$, where $\alpha_l \equiv z_j^{s-1} c_{l-1}$. The argument follows. Finally, it is well known that $\det(V) \neq 0$. Combining these, we see that $\det(\widetilde{M}) \neq 0$. Therefore, $\widetilde{M}$ is non-singular.

Next, we show $A_2$ is non-singular. Note that $\sum_{i=1}^k i^s = \frac{k^{s+1}}{s+1}(1 + o(1))$, $\sum_{i=1}^k i^s \cos^2((i-1)\theta) = \frac{k^{s+1}}{2(s+1)}(1+o(1))$ and $\sum_{i=1}^k i^s \sin^2((i-1)\theta) = \frac{k^{s+1}}{2(s+1)}(1+o(1))$, for $\theta \neq -\frac{\pi}{2}, 0, \frac{\pi}{2}$. Also, $\sum_{i=1}^k i^s \sin((i-1)\theta) = o(k^{s+1})$ for all $\theta$, and $\sum_{i=1}^k i^s \cos((i-1)\theta) = o(k^{s+1})$ for $\theta \neq 0$. Using these arguments and basic equalities in trigonometric functions, we have

$$(R'R)_{(j,s),(j',s')} = o(1) + \begin{cases} \frac{\sqrt{(2s-1)(2s'-1)}}{s+s'-1}, & j = j', \\ 0, & \text{elsewhere.} \end{cases}$$

As a result, $A_2$ is a block-diagonal matrix, where each block corresponds to one $z_j$ on the unit circle and is equal to the matrix $W(h_j)$, where $h_j$ is the replication number of $z_j$ and $W(h)(s,s') = \sqrt{(2s-1)(2s'-1)}/(s+s'-1)$, for $1 \le s, s' \le h$. Since such $W(h)$'s are non-singular, $A_2$ is non-singular.

□

## ACKNOWLEDGEMENTS

The authors would like to thank Ning Hao, Philippe Rambour and David Siegmund for helpful pointers and comments.

## REFERENCES

[1] Andreou, E. and Ghysels, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *J. Appl. Econometrics*, **17**, 579-600.

[2] Bhattacharya, P. K. (1994). Some aspects of change-point analysis. *IMS Lecture Notes Monogr. Ser.*, **23**, 28-56.

[3] Candes, E. and Plan, Y. (2009). Near-ideal model selection by $\ell^1$-minimization. *Ann. Statist.*, **37**, 2145-2177.

[4] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when $p$ is much larger than $n$ (with discussion). *Ann. Statist.*, **35**, 2313-2404.

[5] Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**(1), 33-61.

[6] Chen, W.W., Hurvich, C.M. and Lu. Y. (2006). On the correlation matrix of the discrete Fourier transform and the fast solution of large Toeplitz systems for long-memory time series. *J. Amer. Statist. Assoc.*, **101**(474), 812-822.

[7] Donoho, D. (2006). Compressed sensing. *IEEE Trans. Inform. Theory*, **52**, 1289-1306.

[8] Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci.*, **105**, 14790-14795.

[9] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likehood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1349-1360.

[10] Fan, J. and Lv, J. (2008). Sure independence screening in generalized linear models with NP-dimensionality. *J. Roy. Statist. Soc. B*, **70**, 849-911.

[11] Fan, J. and Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*. Springer, NY.

[12] Friedman, J. H., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

[13] Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**(1), 1-22. http://cran.r-project.org/web/packages/glmnet/index.html.

[14] Frieze, A.M. and Molloy, M. (1999). Splitting an expander graph. *J. Algorithms*, **33**(1), 166-172.

[15] Genovese, C., Jin, J., Wasserman, L, and Yao, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.*, **13**, 2107-2143.

[16] Harchaoui, Z. and Levy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty, *J. Amer. Statist. Assoc.*, **105**(492), 1480-1493.

[17] Ising, E. (1925). A contribution to the theory of ferromagnetism. *Z. Phys*, **31**(1), 253-258.

[18] James, B., James, K.L. and Siegmund, D. (1987). Tests for a change-point. *Biometrika*, **74**, 71-83.

[19] Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high dimensional variable selection, *Ann. Statist.*, **40**(1), 73-103.

[20] Jin, J., Zhang, C.-H. and Zhang, Q. (2012). Optimality of Graphlet Screening in high dimensional variable selection. *arXiv:1204.6452*.

[21] Lehmann, E., Casella, G. (2005). *Theory of point estimation* (2nd ed). Springer, NY.

[22] Meinsausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the the lasso. *Ann. Statist.*, **34**, 1436-1462.

[23] Moulines, E. and Soulier, P. (1999). Broadband log-periodogram regression of time series with long-range dependence. *Ann. Statist.*, **27**(4), 1415-1439.

[24] Niu, Y.S. and Zhang, H. (2011). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.*, to appear.

[25] Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**(4), 557–572.

[26] Rambour, P., Seghier, A. (2005). Szego type trace theorems for the singular case. *Bull. Sci. Math.*, **129**(2), 149-174.

[27] Ray, B.K. and Tsay, R.S. (2000). Long-range dependence in daily stock volatilities. *J. Bus. Econom. Statist.*, **18**(2), 254-262.

[28] Siegmund, D.O. (2011). Personal communication.

[29] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B*, **58**(1), 267-288.

[30] Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9**, 18-29.

[31] Wade, N. (2009). Genes show limited value in predicting diseases. *New York Times April 15*. http://www.nytimes.com/2009/04/16/health/research/16gene.html.

[32] Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.*, **37**(5), 2178-2201.

[33] Yao, Y.-C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhya, A*, **51**, 370-381.

[34] Zhang, N.R., Siegmund, D.O., Ji, H., and Li, J. (2010). Detecting simultaneous change-points in multiple sequences. *Biometrika*, **97**, 631-645.

[35] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Mach. Learn. Res.*, **7**, 2541-2567.

T. Ke and J. Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, New Jersey, 08544
USA
E-mail: zke@princeton.edu
jqfan@princeton.edu

J. Jin
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213
USA
E-mail: jiashun@stat.cmu.edu